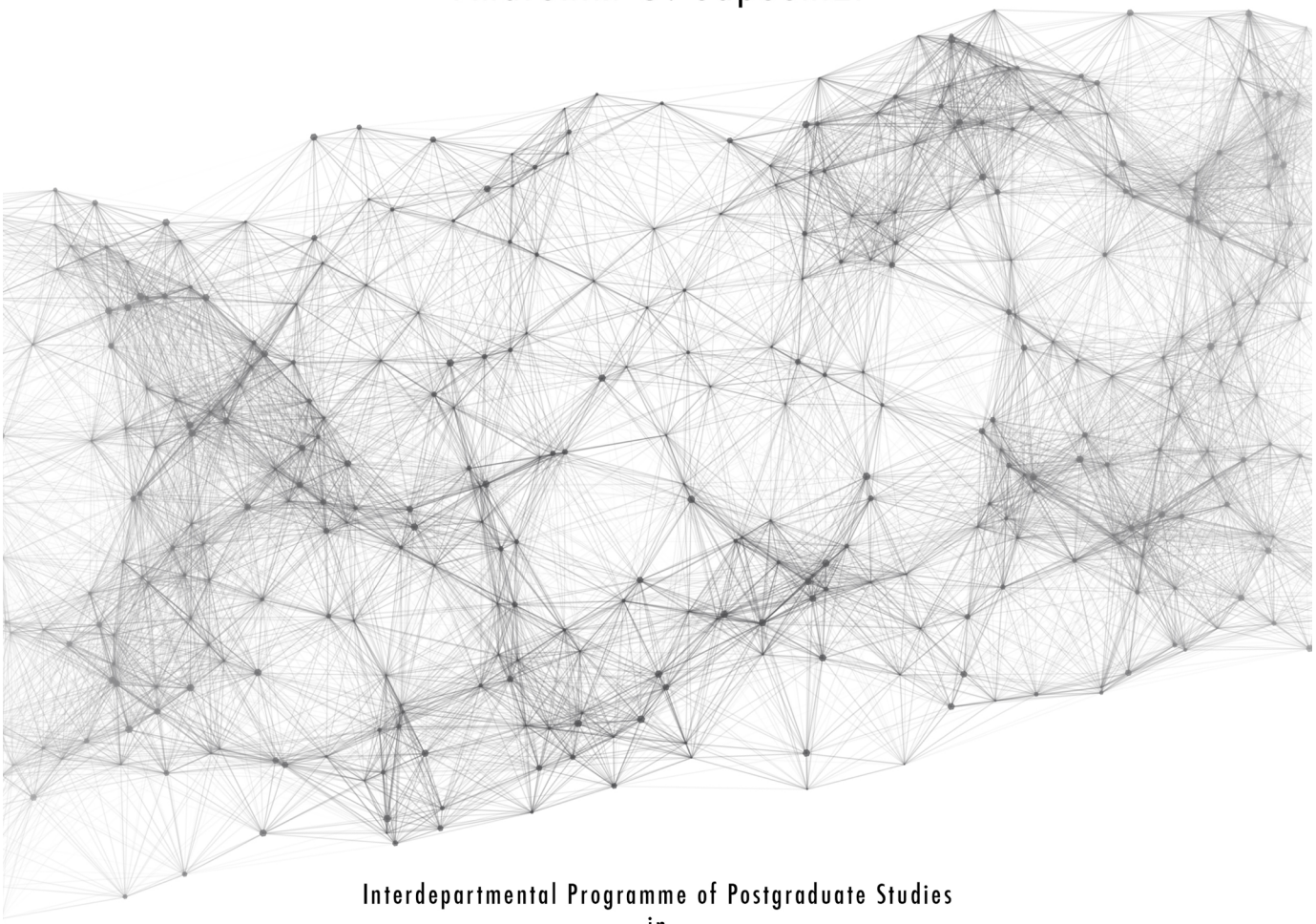
HELLENIC REPUBLIC
UNIVERSITY OF MACEDONIA

MSc Dissertation

Insights from Social Networks: A Big Data Analytics Approach

by

Androniki G. Sapountzi



Interdepartmental Programme of Postgraduate Studies
in
Information Systems

Submitted as a prerequisite in fulfillment of the requirements for the acquisition of the postgraduate degree in Information Systems

10/2016
Thessaloniki, Greece.

ACKNOWLEDGEMENTS

This research is the final result of my Master Thesis project to obtain the Master degree of Information Systems at the University of Macedonia in Thessaloniki. I would like to thank several people who supported me during this Master Thesis project.

First and foremost, I wish to thank my thesis advisor, Psannis E. Konstantinos, professor of the department of Applied Informatics. He helped me come up with the thesis topic and guided me over almost a year of development with his vast knowledge and skill. During the most difficult times when writing this thesis, he gave me the moral support and the freedom I needed to move on. Secondly, I would like to thank both my colleagues and professors from the university with whom I made constructive conversations that rendered many sparkling ideas to be materialized during my research. A very special thanks goes out to my friends, without whose motivation, love and encouragement, I would not have finished this thesis. Last but definitely not least, I would like to thank my parents and my two brothers for their continuous support during my study.

Androniki Sapountzi

Thessaloniki, October 2016

ABSTRACT

One of the biggest domains of insights of Big Data are online social networks(OSN), whose paths for knowledge are currently under exploration. The unfolding of every event, breaking new or trend flows in real time inside OSN triggering a surge of opinionated networked content. Such unprecedented scale of human communication and public behavior data brings new opportunities to understand how society works. Tools are fundamental to help people perform data analysis tasks. However, meager progress is done; not only because science is still far from automatically processing human-centric data but contextual concerns such as privacy or noise restricts this endeavor as well. OSNs aren't yet examined at the cleansing stage and until now data analysis has been studied separately. With this research we are trying to make the start. Investigating the social networks as a contextual source, their data value chain, the low quality data they contain and how could these be addressed, constitutes the first step. Secondly, the entire spectrum of social networking data analysis, namely (i) social network analysis, (ii) sentiment analysis, (iii) topic detection and (iv)collaborative recommendation is studied. In particular, our purpose is to exploit state-of the-art frameworks and techniques and correlate them in both cleansing and analysis services. Then, we develop a cross-platform tool for sentiment analysis through modern cloud-hosted machine learning services. Lastly, the idea is to capture both analysis limitations and future trends regarding data from OSN with a special interest on sentiment analysis and computational intelligence paradigm.

Περιληψη

Μια από τις σπουδαιότερες πηγές μεγαδεδομένων είναι τα κοινωνικά δίκτυα (ΚΝ), των οποίων τα μονοπάτια εξόρυξής γνώσεων είναι υπό εξερεύνηση. Κάθε τάση, συμβάν ή είδηση ξεδιπλώνεται στον χώρο των ΚΝ σε πραγματικό χρόνο προκαλώντας χειμαρρώδη ροή δικτυωμένων απόψεων. Ιδιαίτερα πρωτοφανές είναι το μέγεθος των δημόσιων δεδομένων κοινωνικής συμπεριφοράς και επικοινωνίας το οποίο ανοίγει δρόμους ερμηνείας για τον τρόπο λειτουργίας της κοινωνίας. Τα εργαλεία είναι θεμελιώδη στην προσπάθεια των ανθρώπων να απαντήσουν ερευνητικά ερωτήματα μέσω της ανάλυσης δεδομένων. Ωστόσο πενιχρή πρόοδος έχει διατυπωθεί, αφενός διότι ο επιστημονικός τομέας αναλυτικής κοινωνικών μεγαδεδομένων βρίσκεται σε νηπιακή ηλικία και αφετέρου ανησυχίες σχετιζόμενές με το πλαίσιο των ΚΝ όπως η ποιότητά και η ιδιωτικότητα των δεδομένων περιστέλλουν εξίσου το παρόν εγχείρημα. Τα κοινωνικά δίκτυα δεν έχουν ερευνηθεί σε επίπεδο επεξεργασίας-καθαρισμού δεδομένων καθώς έως τώρα ιδιαίτερη σημασία τοποθετήθηκε στο στάδιο της ανάλυσης το οποίο μελετιόνταν ξεχωριστά από αυτό της επεξεργασίας. Αντικείμενο της παρούσας έρευνας αποτελεί αρχικά η διερεύνηση του πεδίου δεδομένων των κοινωνικών δικτύων, της αλυσίδα αξίας των δεδομένων, τα είδη χαμηλής ποιότητας δεδομένων καθώς και η διαχείριση τους. Εν συνεχεία, διερευνάται ολόκληρο το φάσμα των εδραιωμένων πρακτικών ανάλυσης δεδομένων των ΚΝ, ήτοι (i)ανάλυση κοινωνικών δικτύων, (ii) ανάλυση συναισθημάτων, (iii) ανίχνευση συγκεκριμένων θεμάτων και (iv) συνεργατικά συστήματα συστάσεων. Ειδικότερα, σύγχρονα εργαλεία και επιστημονικές τεχνικές διερευνώνται και συσχετίζονται με υπηρεσίες καθαρισμού και ανάλυσης. Έπειτα προγραμματίζεται ένα ανεξάρτητο τεχνολογίας εργαλείο το οποίο εκτελεί ανάλυση συναισθημάτων διαμέσου αυτοματοποιημένων μηχανικής μάθησης μοντέλων βασισμένων στο νέφος. Τέλος, διατυπώνονται οι περιορισμοί των τεχνικών αναλύσεων και οι μελλοντικές τάσεις, με ιδιαίτερο ενδιαφέρον προς την υπολογιστική νοημοσύνη και την ανάλυση συναισθημάτων.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 MOTIVATION AND PROBLEM STATEMENT

Our networked world with the ubiquitous data creation arose the entering in a new scientific age, called the 4th industrial revolution as predicted by Jim Gray (2009) and has recently attracted public attention (Parker and Thomson(2016), Djorgovski(2015), Degryse(2016), Lesk(2016), Hey(2012), Hannay(2015), Markoff(2009)). It describes the rapid change of science as a result of the collection and analysis of the vast amounts of data, also called Big Data.

Big Data is considered a powerful tool that reframes many key questions such as the constitution of knowledge (Boyd and Crawford (2012)) and is widely characterized by the 3V model (Laney 2001): Volume of data, Variety of data types and Velocity at which data processed. IBM transform it to 4V by adding the attribute Veracity (Gandomi and Haider(2015)) which came as a response to the weakness of dirty data and it is one of the most widespread definitions of what is known as Big Data Problem (Saha and Srivastava (2014)). All V's are directly connected with data analysis. Still, both scientific and social issues are involved in Big Data analysis.
Three of the most common big data sources are the Internet of Things (IoT), Cloud Computing and Online Social Networks(OSN) (Meng and Ci (2013)). Figure 1 depicts a classification of big data, carried out by Pattnaik and Mishra (2016). This research focus on data generated in OSN's since they are considered one of the biggest domains of social insights.

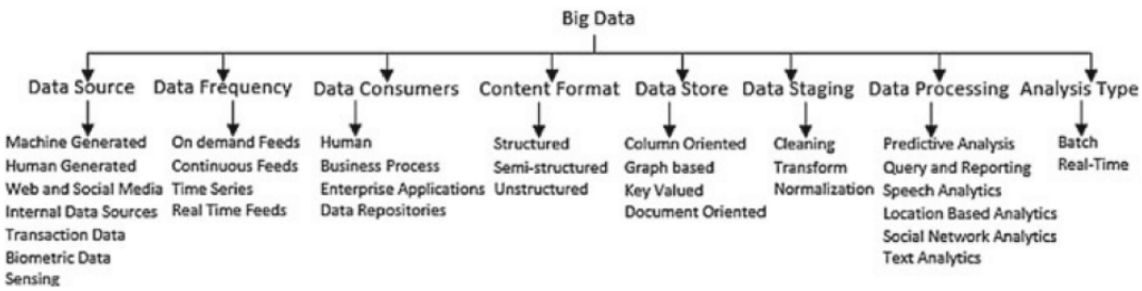


Figure 1: Classification of Big Data

The advent of mass adoption of online social networking sites has caused a shift on how

people communicate and share knowledge, how businesses operate and compete and how government act and influence. In the research area, it has almost replaced any conventional social science tool (interviews, questionnaires) announcing thus the computational social science (Kaisler et.al (2013); Shin and Choi (2015); Mackey (2013)). The impressive growth of social networking services (SNS) makes available an unprecedented scale of personal data, data about events and social relationships, public sentiments and behaviors that when are mined and interpreted are of an enormous value. Figure 2 highlights the statistics of Internet users among online communities (Chaudhary et.al (2016)).

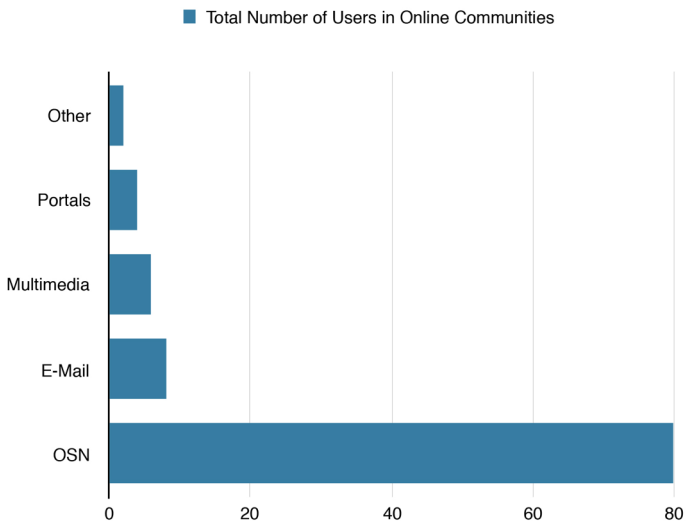


Figure 2: Popularity of OSN among online users

Although many scientific endeavors have been done, deriving knowledge from social network-sourced Big Data still remains a challenge principally because of two reasons.

Firstly, the social nature of nodes in social networks makes data subjective to many privacy concerns. Actually, the biggest challenge of Big Data is indeed privacy (Pentland (2016); Mo & Li (2015); Hewlett Packard(2015)) and deficit to say that, when the source is social networks all challenges related to Big data become even more salient.

Secondly, science is still far from automatically analyzing unstructured human communication data because machines are not yet able to understand human language; and therefore social big data science is still developing. Additionally, the garbage input garbage output adage of yore is alive and well. Due to the the informal language data exchange over OSN and the medium's noisy nature, conventional technologies of

preprocessing are inadequate. However, to the best of our knowledge, current data cleaning techniques are spread across different domains apart from that of social networks.

The research is guided toward the second challenge.

1.2 Research Goal

Social network analysis, Topic detection and Tracking, Sentiment Analysis and Collaborative Recommendation are the four prevalent analysis practices in OSN. It is clear that analytics is a complex process that demands people with expertise in cleaning up data, understanding and selecting proper methods, and analyzing results. Tools are fundamental to help people perform these tasks. However, even more tangled the knowledge discovery process has become with the arrival of big data era and new tools are constantly arise to replace the conventional no-effective ones. Regarding the area of social networking there is much confusion among data scientists due to the lack of standardization of processes and data quality.

In response to this chaotic emerging science of social data and predictive knowledge, the main goal of this thesis is to contribute to the above knowledge gap by approaching social networks through big data analytics and specifically from the technical perspective of tools, techniques and services.

The broad research question that fit this goal is:

> "What are the recent frameworks' services regarding social networking data analysis, how are they developed and how can the involved processes be optimized?"

One way to get more value out of the available data and optimize the analysis process is to have 'well-cooked' data, since analysis results highly depend on the quality of data. Therefore, Veracity and data cleansing are also scrutinized. In this end, up-to-date data analysis and cleansing frameworks of the field are surveyed, considering the different kinds of analysis, the diversity of methods and the functionalities offered by these tools. Since the inherent characteristics of CI algorithms are of paramount importance in addressing big data analytics' critical issues, a correlation between CI techniques and OSN analysis is also presented.

1.3 Thesis Overview

The rest of the paper is structured as follows. In Section 2, the related work together with the theoretical framework is presented. An introduction about the categories of data types and data analysis methods and practices is also discussed. In section 3 the steps of research methodology that fulfill the thesis' goal is given. Section 4 tackles the problem of data preprocessing and cleaning in social networks. In section 5 a comparative analysis among social network analysis tools and a correlation between tools' inherent metrics and graph-analysis methods is given. Section 6 and 7 provide a sophisticated classification among trendy topic detection and sentiment analysis frameworks respectively. Collaborative recommendation frameworks, including their related algorithms and techniques, are investigated in Section 8. In section 9 the analysis issues of the common data analysis approaches and the potentials of Sentic Computing and Computing Intelligence paradigm is discussed. Section 10 illustrates the set up of the experimental environment. A cross platform tool for sentiment analysis via machine learning APIs is developed and employed with Indico and Google Natural Language (NL) API on public classified opinionated datasets in an attempt to evaluate the current state of sentiment analysis effectiveness in OSN. The conclusion and possible future work are demonstrated in Section 11.

2 LITERATURE REVIEW

In order to derive knowledge gaps, a theoretical framework of the main research achievements in the science of social networking data cleansing and analysis is established with this chapter. Simultaneously, asking questions related to these gaps, creates a flow of important substances beneficial to contributing in the research community.

2.1 USED METHOD FOR LITERATURE REVIEWING

A vital step to create a proper foundation for any researcher is the task of completely reviewing a chunk of academic literature (Wolfswinkel et al., 2013). Considering the vastness of papers and literature, criteria for inclusion or exclusion of resources are essentially to be defined. Initially, papers that are both in the scope of this study and related subject areas, without overlooking the credibility of the publication sources, are searched. Based on the dimensions of information quality (Taleb et.al (2015); Lee et.al (2002)) we select recent papers, with a special interest on those published within the period of 2014-2016, that are mostly cited, considering concurrently the analogy between time and number of citations. We could say that this step is similar to big data cleaning (Boyd & Crawford (2012)) since both borrows the principal of gardening as wisely Sankaranarayanan et.al (2009) pointed:

"if you are careful about the seeds that you plant,
you will only grow the plants that you desire".

2.2 ONLINE SOCIAL NETWORK (OSN)

Network concepts and techniques are widely found throughout a range of disciplines; the entire world around us poses a network structure. Economy, human cell, traffic and roads, society, internet, food webs, media and information all have the structure of a network which is commonly modeled by a graph.

Social network is a term used to describe web-based services that allow individuals to create a public/semi-public profile within a domain such that they can communicatively connect with other users within the network. In the most basic framework the social

network is represented as a graph $G = (V, E)$ where V is a set of nodes and E is a set of edges that connects the nodes.

The study of social networks is a new but quickly widening multidisciplinary area involving social, mathematical, statistical, and computer sciences. The unique element of social networked data is that they reveal information about interactions between users-communities-content. However, each social network views their users through radically different lens and have no same network representation making it difficult to integrate data from different social networks. A reason justifies that is that, each social networking service (SNS) provides a platform that attracts people to built a specific type of networking relationship range from professional (LinkedIn) to research (Research Gate). From a bird's eye view, Facebook has interest graph whereas Google advocates knowledge graph and regarding the context of language exchange Twitter is considered as informal whereas LinkedIn as formal.

The mass adoption of SNS is considered as a spark that burst the Big Data era and arise great opportunities to the understanding of most socio-economic phenomena in the modern world. OSN is a rich source of opinionated text and multimedia content that have recently gained huge popularity especially in the area of political and marketing campaigns. Social network information also has recently incorporated in recommendation systems. The latter are capable of dealing with the problems of information overload and information filtering.

Social networks have transform many aspects of our daily lives. To understand that impact, look no further than how the movie rental experience has changed which has become a service that utilizes a vast array of data points to generate recommendations (EY (2014)). The diffusion of breaking news, especially in Twitter, is considered to be disseminating much faster than in any conventional news media. In this end, early event detection and social network analysis play a detrimental role in management of natural disasters, epidemics and terrorism breakouts. Businesses also apply social network analysis to gain insight into markets and communities (Hansen et.al (2010)), with the "social enterprise" being the new necessity in order to manage knowledge and cooperation. Alex Pentland and Asu Ozdaglar have recently created the MIT Center for Connection Science and Engineering for understanding connections like how people are connected together by machines and how, as a whole, they create a financial market, a government, a company, and other social structures.

Having shifted away from the analysis of single small graphs and the properties of individual nodes to consideration of large-scale properties of graphs, the need for new data analysis tools and techniques is arisen.

2.3 Big Data Quality & Cleaning

The prevalent attributes of data incompleteness, inconsistencies, unreliability and timeliness in social networks hampers all the stages of network data analytics and affect data quality dimensions. Preparation purpose is to both revert the data to a format capable for the analysis process and to ensure the high quality of data; it consists of (Vaidya (2016)) the following techniques:

1. Data Cleaning/Cleansing/Scrumming,
2. Data Integration,
3. Data Transformation,
4. Data Reduction and
5. Data Summarization.

Many researchers noticed that quality standards for big data are missing and consequently proposed big data quality frameworks. Saha & Srivastava (2014) provide a big data quality management view that corrects data via logical/constraint model, based on rules which should be learned by the data itself (auto-discovery). Taleb et.al (2015) propose a big data pre-processing system that aims to manage data quality at all processes with the selection of rules to be user-defined, auto-discovery or domain related. They divided quality dimensions, as illustrated in Figure 3, into the intrinsic dimensions, that refer to objective data attributes, and the contextual ones. Unece HLG Big Data Project (2014) present a theoretical big data quality framework that uses a hierarchical structure composed of three hyper dimensions: the source, the metadata and the data, with quality dimensions nested within each hyper dimension. Cai and Zhu (2015) also provide a hierarchical structure of a data quality framework composed of data quality dimensions accompanied it with a dynamic quality assessment process. Immonen et.al (2015) present a big data processing architecture to manage the quality of social media data by utilizing enterprise's data policy rules and metadata creation and provenance. A timeless work has been done by Rahm and Do (2000) who classify the data quality problems into two categories: *single-source* and *multiple-source* problems that are further divided into *schema* and *instance* levels. Although this framework is not specific to social networks, the issues in data

cleaning for social network analysis can be clearly identified from their perspective according to Bonchi et.al (2011).

Particularly in social networks the collection process relies on samples obtained via stream APIs without consideration on the quality of the samples. How to define "on-line" filters in such a way that they do not discard useful information is still a challenge (Jagadish et.al (2014)). All of the researchers argue that data cleaning is both one of the perennial challenges in big data analytics and critical to knowledge discovery. With the advent of big data, data quality management has become more important than ever and new challenges have emerged such as the need for context-aware data quality rules and the fast and scalable algorithms to ensure data quality.

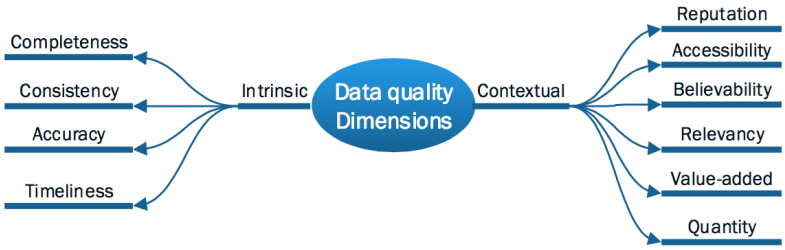


Figure 3: Data Quality Dimensions

There is an increasing interest both from academia and industry in data cleaning and transforming tools. Traditional data cleaning tools include ETL, Excel spreadsheets and Google Refine (Batrinca & Treleaven(2015)). Pulla et.al (2016) compared state of the art open source data quality tools. They conclude that the most efficient of all is DataCleaner which has recently take the advantage of cloud and the big data technology of Hadoop to analyze the state of the data. KNIME is another modern data cleaning tool used in handling big data and social media data (Chen et.al (2014); Minanovic et.al(2014)). Two other tools for big data cleaning are Data Wrangler (Kandel et.al (2011), oriented toward individual data scientists and similar to Google Refine, and Data Tamer, enterprise-oriented.

Nevertheless, there is still a demand for progress. Maletic and Marcus (2009) stated that most cleaning tools address the duplicate detection problem while Shuguang et.al(2016) pointed that automated techniques and tools for streamlining the social media analytics process are still missing.

A taxonomy of data cleaning techniques is provided by Chu et.al (2016). Their work can be considered as an endeavor to approach the issue of the absent framework for

classification of data cleaning problems (Hu et.al (2014)). They regarded qualitative data cleaning techniques which encompasses constraints, rules, or patterns to detect and repair errors and the quantitative perspective which employs statistical and Machine learning methods. Chu et.al (2016) and Saha & Srivastava (2016) among other researchers consider qualitative methods to outperform quantitative ones. Xu (2016) compared current quantitative data cleaning methods both the traditional applied statistics and Machine Learning techniques for overcoming big data challenges and generally argued that there is no universally applicable cleaning method.

Few literature deals with unstructured data cleaning, though recognize it as an open research topic. Maletic and Marcus (2010) have noticed that in the area of social networks many statistics are published without the explanation of how data collected, cleaned and analyzed, leaving readers unable to assess their results. Neither a holistic comparison of data cleaning techniques nor cleansing social network data is published yet.

Importantly, cleansing dirty data is a hard task and studied for decades, since not only is itself prone to errors (Khayyat et.al (2015)) but also data errors arise in different forms. Understanding these sources of error is a first step toward developing a data cleaning technique (Jagadish et.al (2014)). To achieve that, an investigation of what data types are generated in social networks and what are the popular analysis algorithms that such data are used, is required.

2.4 DATA FROM OSN

Social networks typically contain a tremendous amount of *content* and *linkage* data which can be leveraged for analysis. The linkage data is essentially about patterns of interactions between the network entities (e.g., people, organizations, and products) and measured by Social Network Analysis(SNA); whereas the content data is User Generated Content(UGC), the lifeblood of SNS, and includes text, images, videos, tweets, product reviews and other multimedia data created and shared in the network, typically studied with content-based analysis (Gandomi & Haider (2015); Aggarwal (2011)). These types can further be divided into unstructured and structured data respectively depending on whether they are organized in a pre-defined manner (structured data) or not (unstructured data). To illustrate this with an example, time-based events are structured, whereas event data based on tweets and “likes” are unstructured. Structured data in OSN are usually

graph-structured. In the most basic framework, they are modelled with a social network which is represented as a graph**.** To illustrate the difference between the two, time-based events can be considered structured, whereas trend data based on tweets, re-tweets and "likes" are unstructured. Figure 4 summarizes the types of data and the corresponding analysis conducted in OSN.

In general, the variety of data collected and leveraged for analysis can be distinguished in *explicit data*, i.e. information directly related to service usage (e.g. profile details, interests, number of friends, etc.), and *implicit data*, i.e., that are either information that is processed automatically in the system (e.g. browser data, web sites visited, etc.) or can be discovered from user's activities by analyzing extensive and repeated interactions between users (voting, sharing, tagging, commenting items) (Bonchi et.al (2011); Chen et.al(2016)). Researchers like Chen et.al (2016)) leverage implicit data for personalized recommendation in Twitter whilst He and Chu (2010) are interested in explicit and other work like Hu et.al (2013) focus on both implicit and explicit data.

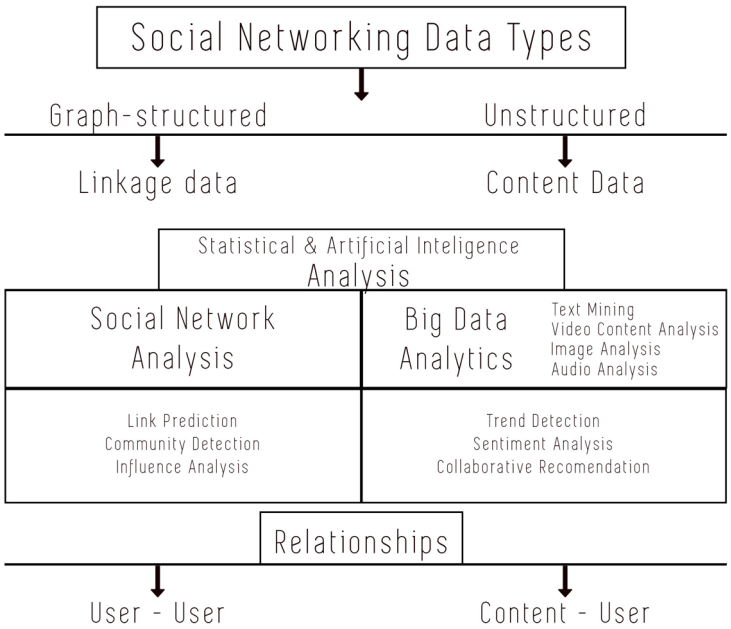


Figure 4: Data Types and Analysis

Social network analytics and content mining approaches follow the interdisciplinary principles of Artificial Intelligence (AI), Statistics and related areas. Decades before the advent of OSN AI researches attempted to embed the controversial notion of 'intelligence' in machines so as to comprehend, reason and learn about how the world works and hence acquire further capabilities from mere logical computations (Siddique and Adeli (2013); Cambria and Hussain (2015)). OSN can be used as an environment of

endowing machines with the capacity of this common-sense knowledge. The last few years have seen rapid progress on long-standing, difficult problems in AI and it is now rapidly reinventing so many of the Internet's most popular services (Metz (2016); Davis and Marcus (2015); Clark (2016); Amodei et.al.(2016)). Statistics on the other hand involve less intricate procedures that emphasize to statistical models towards the better understanding of data generating process.

Public APIs are the standard mean of retrieving social networking data from cloud and they typically encourage the development of third-party software—for example, a plugin for WordPress. One alternative is to use commercial tools for scrapping that protect raw data or that have some extra filtering functionality. For instance, Kaushik et.al (2016) used Sysomos, a social monitoring tool, to detect specific events. Sysomos is also one of the tools used at the BBC for monitoring social media and website activities (MacKay 2013). Another alternative is to use the combination of API and a crawler as (Cagliero and A. Fiori (2013)) did. A crawler is built to extract information that are not automated to be extracted with service API. Importantly, though, each social platform has very specific rules around how on to use their respective data that can be found in the Terms of Service. Although, most of SNS expose an Application Provider Interface (API) which includes methods to get a range of data including friends, events, groups they limit the number of API transaction per day.

2.5 Data Science in OSN

Recent developments in technology such as cloud computing and big data analytics advocate the mining of insights in OSN. Social media sites have a large number of user scattered across the globe which makes them ideal candidates for cloud adaptation. Big data analytics are being applied in social networks to extract meaningful insights through text mining and multimedia mining (Gandomi & Haider (2015; Tanwar et.al (2015)). An open issue in big data analytics according to a recent survey (Tsai et.al (2015) is the usage of soft computing algorithms since, although they can analyze such complex nature of data, unfortunately, until now, not many studies are focused on it. Soft computing is the basis of Computational Intelligence which in contrary to AI-based systems, does not require the construction of precise models to deal with the imprecise, incomplete, and uncertain information (Siddique and Adeli (2013)).

SNA is important if one wants to understand the structure of the network so as to gain

insights about how the network "works" and make decisions upon it by either examining node/link characteristics (e.g. centrality) or by looking metrics at the whole network cohesion (e.g. density) (Hansen et.al (2010); David & Jon (2010); Kolaczyk and Csárdi (2014)). For instance, some indicators that measure the influence and credibility of a user are mention influence, follow influence, and retweet influence and can measured through centrality analytics. Graph theory is the core prominent approach in social network analysis and used to investigate social structures both analytically and visually. One of the main obstacles regarding SNA is the vastness of Big Data since analysis of a network consisting of millions or billions of connected objects is usually computationally costly (Chen et.al (2014)).

Content analysis studies unstructured content generated in OSN's by users while a lot of interest has been placed in extracting meanings from the textual data through text mining techniques. Analysis practices in OSN's include the following: (Gandomi and Haider(2015); Thiel et.al (2012))

- Topic Detection and Tracking(TDT),
- Sentiment Analysis (SA) and
- Collaborative Recommendation (CR).

Sentiment Analysis is an ongoing field of research in text mining that determines people's opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics and their attributes (Liu and Zhang(2012)). TDT is about discovering the emergence of new topics (or events) and tracking their subsequent evolvements over a period of time (Adedoyin-Olowe et.al (2013)). It requires the automatic answering of "What, when, where and by whom are the popular topics/trends being set" but until now, it is clear that no method addressed all of these questions (Panagiotou et.al (2016)) efficiently. Collaborative recommendation refers to Collaborative Filtering approach that predicts a target user's interest in particular items based on the opinions or preferences of other users (Sieg et.al (2010)).

There are two ways to conduct an analysis in OSN:

(i) *Static or batch* analysis presumes that social network changes gradually over time.

(ii) *Dynamic* analysis, which is more intricate, encompasses streaming data that are evolving in time at high rate. It is often in the area of interactions between

entities whereas static analysis deals with properties like connectivity, density, degree, diameter and geodesic distance.

Social network analytics and content mining are not mutually exclusive, far from it, should co-exist in analysis. Content information in different parts of the network is often closely related to its structure (Aggarwal (2011)) and therefore combining both two sources of information is useful in discovering of hidden patterns. For instance, sentiment analysis can use both linkage data and unstructured text. Previous sentiment analysis approaches often assumed that texts are independent; but in the context of Social Networks, data are networked and this feature shouldn't be overlooked (Hu et.al(2013)). In addition to that, social relationships among users are valuable information in recommender systems but they should also should include what content is shared among users. In Table 1 we illustrate analysis types in OSN that has gained remarkable attention both from academic and marketplace community. The colors indicate the types of analysis usually employed together.

Table 1: Questions handled by popular analyses in OSN.

Analysis Type	Question Handling
Topic Detection and Tracking	"What, where. when and by whom set the popular topics?"
Trend Analysis	"What will become trend and fashion?"
Influence Analysis (Centrality Analystics)	"Who are the popular users in the network?"
Community Detection and Analysis	"Where the network splits up into smaller groups that share a common pattern and what pattern is that?"
Collaborative Recommendation	"What is recommended for a user in relation to the network they belong"
Web Analytics	"What is the performance of a specific site, based on the behaviour of its users"
Link analysis	"What are the inferred pattern of relationships between users?"
Path Analysis	"What is the best possible path of edges to reach a certain node?"
Sentiment Analysis	"Is any emotion expressed in the content and what is that?"
Opinion Mining	"What other people think toward entities, people, ideas?"
Named Entity Extraction	"What are the entities inferred and in which category belong?"
Relationship Extraction	"What are the semantic relationships between entities?"
Semantic Analysis	"How to represent and infer common sense knowledge from digital content?"
Syntax Analysis(Parsing)	"What are the rules of a formal grammar that conform a string of symbols, either in natural language or in computer languages, conforms to?"

2.5.1 NATURAL LANGUAGE PROCESSING

Content analysis draws on techniques from various fields such as artificial intelligence, data mining, information retrieval and text mining. Generally, commonly used approaches in content analysis can be divided into linguistic, semantic, statistical and/or a combination of them.

The most basic unit of linguistic structure appears to be the word; and fundamental to content analysis operations ranging from training a machine learning model, scoring documents on a query, content classification and content clustering (Manning et.al (2009); Peled et.al (2014)) is the representation of a set of documents as vectors of words, known as the *vector space model*.

Language models are typically used to rank sentences and to compute relevance based on content information. They are trained through a set of string features such as phonemes, letters, or words. Language modelling is a function that puts a probability measure over strings drawn from some vocabulary (Manning et.al(2009)). That is, for a language model M over an alphabet Σ:

$$\sum_{s \in \Sigma^*} P(s) = 1$$

N-gram language model is a contiguous sequence from a sequence of *n* strings of text or speech and when it is of size 1 is referred to as a "unigram", size 2 is a “bigram”, as depicted in Table 2.

Table 2: N-gram model explained through language units

Unit	Sample Sequence	Unigram BoW	Bigram BoW
Word	…As knowledge increases wonder…	…As, knowledge, increases, wonder,…	…As knowledge, increases wonder,…
Character	…to_be_or_not_to_be…	…, t, o, _, b, e, _, o, r, _, n, o, t, _, t, o, _, b, e, …	…, to, o_, _b, be, e_, _o, or, r_, _n, no, ot, t_, _t, to, o_, _b, be, …

N-gram model (Bag of N grams), also known as Bag of Words (BoW), is associated with the statistical measure of the Term Frequency-Inverse Document Frequency (TF-IDF).

However, according to Cambria and White (2014), NLP systems will gradually stop relying too much on word-based techniques while starting to exploit semantics more consistently in order to overcome problems as word-sense disambiguation. The researchers also illustrate NLP research movement in Figure 5 where in the bag-of-

narratives model, each piece of text will be represented by interconnected episodes, leading to a more detailed level of text comprehension solving issues of co-reference resolution and textual entailment. Semantic technologies and NLP have been widely used in many content-based analysis methods both for analysis and cleansing. To incorporate semantic relationships among terms in a vector space model or to retrieve only the relevant information, dictionaries with synonymous such as WordNet have been found useful.

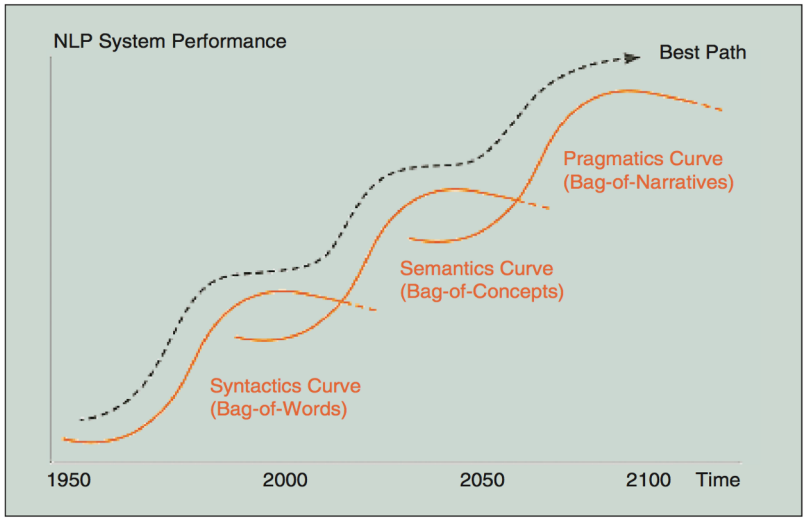


Figure 5: Envisioned Revolution of NLP research

WordNet covers semantic and lexical relations between terms and their meaning such as synonymy, hyponymy and polysemy. Prom on et.al (2015) employed WordNet to find synonyms to expand their manually built subjective set of words in order to analyze sentiments of microblog posts. Ritter et.al (2012) recognized events with the support of dictionaries of event terms gathered from WordNet. Kontopoulos et.al(2013) used WordNet to augment the underlying semantics of the taxonomy of concepts and attributes with synonyms and hyponyms. Also WordNet can be applied to aggregation functions based on hierarchical models where the lower level (e.g. GPS coordinates) features could be aggregated to the higher level (e.g. cities). Additionally, WordNet has been used in a searchable encryption scheme to support personalized search through user interest models (Fu et.al (2016)).

Syntax analysis extract tokes and involves advanced analysis of sentences, terms and term order. It identifies Part of Speech (POS) and Named Entity Recognition (NER) to create dependency parse trees for each sentence, as illustrated in Figure 6. POS and NER methods use sentence structure and language features learned from a large corpus of

annotated text. Another linguistic NLP approach is to perform similarity measurement between clustered noun phrases. Using a graph representation of named entities of the document sets which are connected by dependency relations it may be a good sentence analysis.

On the other side, the TF-IDF measures the significance of words from text ignoring sentence structure. It is a cosine similarity (COS) metric that is used in content analysis usually to score the significance of a word. TF represents the importance of the term within a document and IDF indicates the importance or degree of distinction within the whole document collection. Documents are represented in a Vector Space Model where each document d is represented by the TF vector. TF is the occurrence of the term appearing in the document:

$$dt_f = (tf_1, tf_2, tf_3, \ldots tf_n)$$

where tf_i is the frequency of the ith term of the document d.

IDF gives higher weight to terms that only occur in a few documents and it is defined as the fraction:

$$N/df_i$$

where N is the total number of documents in the collection and df_i is the number of documents in which term i occurs. Another statistical approach is to use heuristic rules, though it is less used.

Some of TF-IDF applications in social media analytics frameworks are listed: calculating similarity between question and topic (Mithun (2013)), training machine learning algorithms (Panagiotou et.al (2016)), retrieving relevant information(Li and Li (2013))

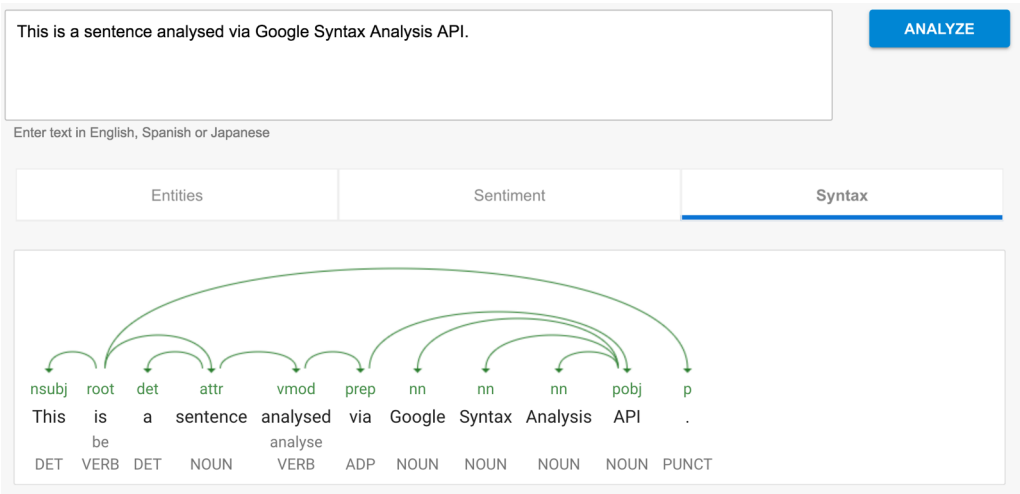


Figure 6: Dependency parsed tree generated via Google NLP API

and enabling multikeyword (Fu et.al (2015)) or personalized (Leung (2013)) ranked search in searchable encryption schemes.

2.5.2 Topic Detection and Tracking

TDT usually employed for detection of emergent or suspicious behavior in the network or for a better understanding of societal concerns (Vakali et.al (2012)). Trend detection is a highly related task to TDT and is commonly applied to social networks.

A useful trend analysis tool that has been used in different disciplines (Yang et.al (2015); Zou et.al (2015)) is Google Trends. As shown in Figure 7, it scores trending topics regarding geographical location and category (e.g. business). Recently, it was found that news topics emerged earlier in Twitter than in Google Trends (Rill et.al (2014)). It is clear that Twitter has become the common place for TDT because it is considered an information network besides a social network (Myers et.al (2014)).
Detecting events relies mostly on machine learning techniques (Atefeh and W. Khreich (2015)). When unspecified events are the case, unsupervised learning is preferred whereas detecting specific events relies on supervised learning. The two main approaches

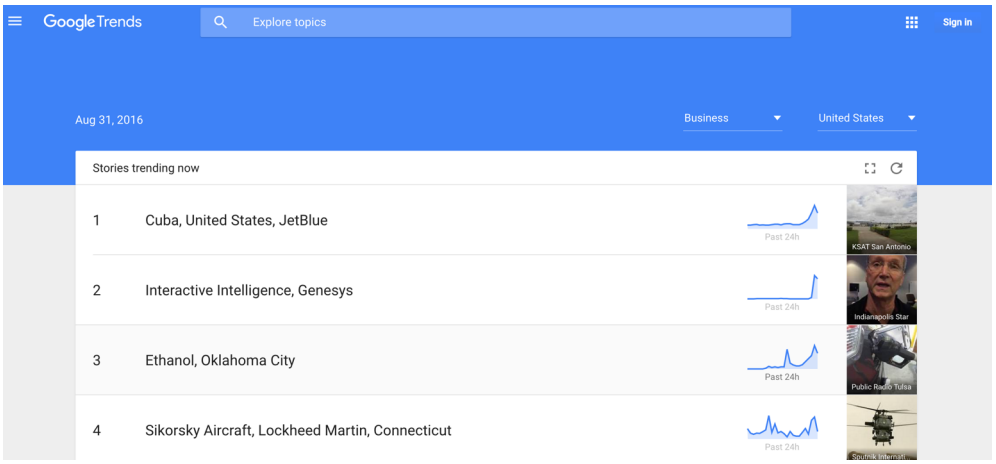


Figure 7: Trend analysis via GoogleTrends

for event detection are classified into

- Feature-Pivot and
- Document-Pivot

depending on whether they rely on temporal or document features (Atefeh and W. Khreich (2015); Panagiotiu et.al (2016)). The former determines trends as those that were previously unseen or growing rapidly and usually focus on burst detection. Twitter presents local trends through this approach, in particular that of term frequency, without providing any additional context for the trending keywords (Vakali et.al (2012)).

The latter is based on textual similarity functions between documents and streams with the support of lexical resources. Both of the two have their limitations. The temporal distributions of features are very noisy and neither all bursts are relevant events of interest (Atefeh and Khreich (2015)) nor all documents are related to events (e.g. memes). Moreover, document-pivot techniques require often batch processing that is not scalable to large amounts of data (Panagiotou et.al (2016)).

A new alternative unsupervised learning technique to the above, is to model normal user behavior and detect any deviation from this baseline profile. It is similar to anomaly detection techniques and has been shown effective in detecting local festival events. Change detection is a common element of TDT; indicators of events considered to be deviation in sentiments, messages' content and the networks' structure (e.g. an increasing number of new connections in the social graph) (Panagiotou et.al (2016)).

2.5.3 Sentiment Analysis

Sentiment analysis refers to detection of the *polarity* as positive or negative in general or about a specific entity. Recently there is an interest, as in "SemEval 2016" is reported (Preslav et.al (2016)), in moving from a categorical two/three-point (plus neutral) scale to an ordered five-point scale namely adding highly positive and highly negative as values, which is now ubiquitous in the corporate world where human ratings are involved: e.g., Amazon, Trip Advisor, and Yelp.

There are two misleading terminologies in SA which are the following:

- Sentiment analysis and opinion mining,
- Polarity and subjectivity classification.

Regarding the first, although opinion mining and sentiment analysis have been used as synonymous terms, the former extracts and analyzes people's opinion about an entity while SA identifies the sentiment expressed in a text then analyzes it (Medhat et.al (2014)). Regarding the second, the basic task of subjectivity classification is classifying a given text into one of two classes: objective or subjective and it is considered as a more difficult task than that of polarity classification whose classification occurs in negative and positive sentiments (Mihalcea et.al (2007)).

Three common approaches to sentiment classification exists in literature, namely,

- statistical which involves mainly machine learning techniques,

- lexicon based methods which leverages affective knowledge bases of words or concepts annotated with their semantic polarity as WordNet Affect, SenticWordNet, SenticNet, AffectiveSpace (Figure 8) and MPQA and
- hybrid approaches which combine the former two.

Dictionaries with synonymous have also been employed to enrich a representation of words like an ontology.

SA techniques can be further divided into three sub-groups namely,

- document-level,
- sentence-level, and
- aspect-level

depending on which textual granularity level will the one sentiment be detected. Classifying text at the document level is mainly based on supervised approaches relying on manually labeled samples of movie or product review data while sentence SA is mainly based either on lexicons by matching the presence of opinion-bearing lexical items (single words or n-grams) so as to detect subjective sentences or on association rule mining for a feature-based analysis of an entity (Poria et.al (2014)). Both of the two do not provide the necessary detail needed opinions on all aspects of the entity. Therefore, we need to go to the aspect level which classifies the sentiment with respect to the specific aspects of entities by firstly identifying the entities and then their aspect (Liu and Zhang (2012)). Again, it is mostly based on supervised machine learning techniques that use language modelling (Schouten and Frasincar (2016)).

After the parable of *USA Today* sentiment analysis about presidential election 2012 (Moore (2012)), where sentiments were derived via mere "word counting" (Struhl (2015)), became apparent that extracting meaningful information from social networks is a much complicated process. There is an endeavor of understanding the natural context with incorporating social networks' structure or with defining the "popularity" and demographic information about the expresser. Smith et.al defined six types of social network structures that tell a story about the nature of the conversation occurring inside OSN, two of them illustrated in Figure 9. Importantly, many social analytics firms like Sysomos, General Sentiment, Crimson Hexagon all have moved to supplement sentiment analysis with other metrics (Kessler (2014)).

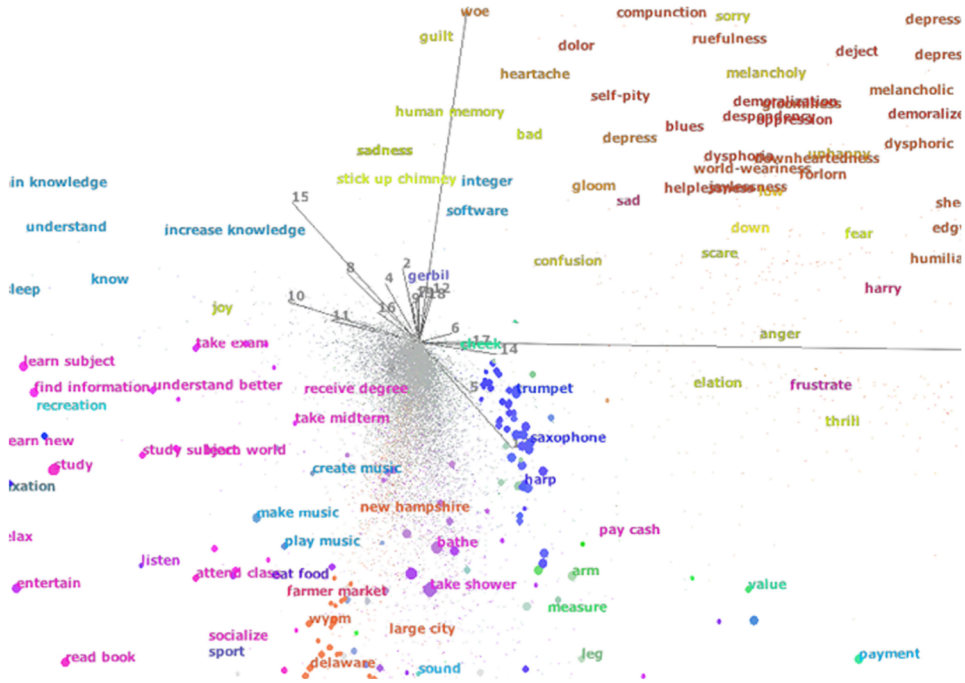


Figure 8: A sketch of Affective Space. Affectively positive concepts are in the bottom-left corner and affectively negative concepts (in the up-right corner). (Poria et.al (2014))

However, research is still needed in the social networks area. Sentiment analysis has been extensively studied for product and movie reviews, which differ substantially from online social networking data (Hu et.al (2013)). In a detailed survey (Medhat et.al(2014)) of the recent adapted approaches related to sentiment analysis was shown that meager academic research upon sentiment analysis has been conducted in the context of social networks. Specifically, from the fifty-four papers only the four were in data scope of OSN's. Also, neither artificial neural networks (ANN) nor fuzzy was used in OSN, which proves the lack of soft computing algorithms usage in big data analytics (Tsai et.al (2015)), as referred in 2.4 section.

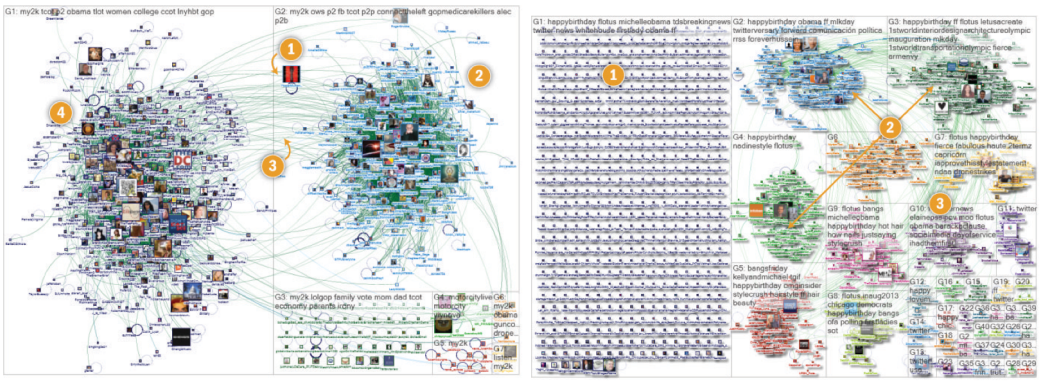


Figure 9: Community Clusters(left) and Polarized Crown Twitter conversation network structures

2.5.4 COLLABORATIVE RECOMMENDATION

Recently, social network information has been utilized as additional input for further improvement of recommender systems, as illustrated in Figures 10 and 11. OSN's permit new forms of rating items, new forms of trustiness and provide user information both at individual and social level. To illustrate this with an example, user generated tags and social relations recently employed by Ma et.al (2015) to augment collaborative recommender systems. However, collecting user interaction data to enhance recommendation accuracy is susceptible to many privacy issues (Atefeh and Khreich (2015)).

Existing recommender schemes can be divided into three categories based on the methods they are build,

- content-based,
- topology-based or collaborative filtering (CF),
- and hybrid approaches that employ both content and topology methods.

The former exploits properties of an item on user past preferences to predict a user's interest towards the item while the second leverages social relations such as user influence and number of common friends and calculates similarities between user profiles to identify users that have relevant interests (Chen et.al (2016); Ricci et.al (2011)). Collaborative recommendation refers to CF approach that determines "What is recommended for a user in relation to the network they belong" by mainly using the feedback from each individual user. There are also variations inside these two approaches. For instance, case-based recommendation system is a variation of content-based approach that recommends items which are similar to what users have indicated as interesting (Ricci et.al (2011)). Community-based system is a variation of CF technique that follows the epigram "Tell me who your friends are, and I will tell you who you are" and it recommends items based on the preferences of the user's friends (Ricci et.al (2011)).

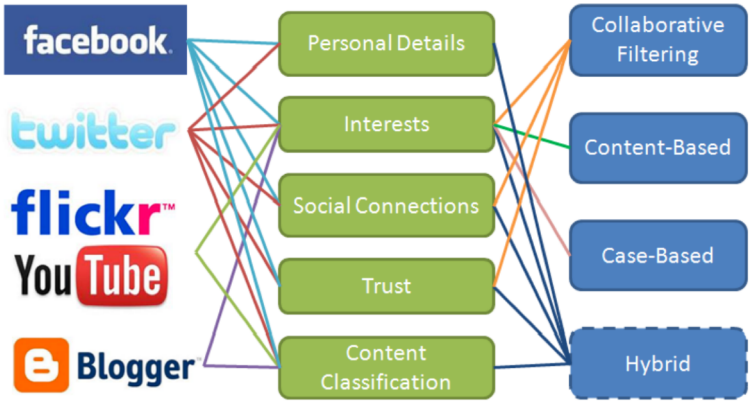


Figure 10: Mapping of Social Web Services and their possible contribution to classical Recommender Systems User Models (Tiroshi et.al (2011)).

CF has emerged as the most prominent approach and is further classified into memory-based (user-based) and model-based (item-based) algorithms. The main idea is that model-based approaches use user-item ratings to learn a predictive model, in contrast, memory-based approaches use user-item ratings stored in the system to directly predict ratings for new items (Yang et.al (2014)). Two of the most popular similarity measurements in selecting potential neighbors are the Pearson Correlation Coefficient (PCC) and Cosine-based Similarity (COS). Though computing PCC or COS for each pair of users can be extremely time-consuming.

Google recently make use of Machine Learning models to provide an API (https://cloud.google.com/prediction/docs/) in order to easily build recommendation systems that are either item-based, user-based or it make recommendations through basket analysis (items frequently bought together).

Item-based techniques enjoy the advantage of easy implementation and avoid the bottleneck of having to search among a large user population of potential neighbor; since first explores the relationships among items (Sieg et.al (2010)).

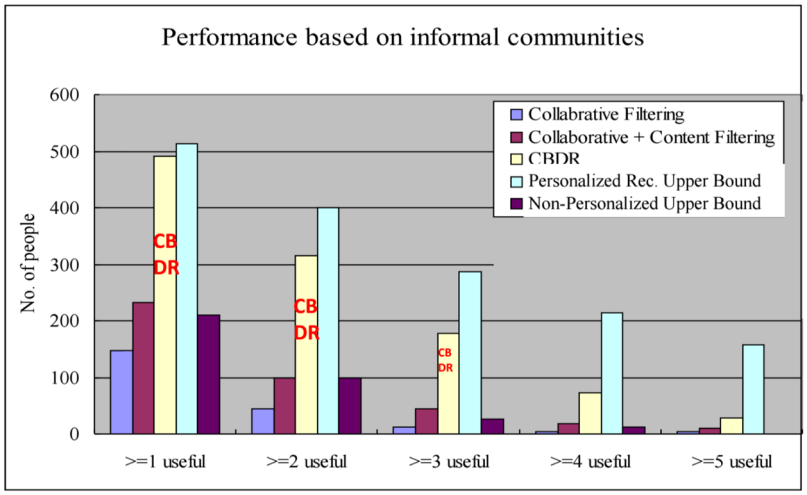


Figure 11: Combining Collaborative and content filtering with graph

Current recommender systems face a lot of issues except for scalability such as data scarcity and the cold star problem and all of them become even more noticeable in the context of OSN. Data scarcity is about limited number of preferences with user/item rating matrix being very sparse. On the other side, cold star problem pertains to the initial membership of a user where no data about their interests are available. Regarding the content relevance calculation, is usually inaccurate due to the short text posts and considering the relevance of a user preference is usually not provided in OSN by explicit features such as user-to-user scores (Chen (2016)).

Another restriction of recommender systems especially those related to user-based method is that they are susceptible to privacy attacks and the violation of sensitive information of users. Privacy-preserving collaborative filtering (PPCF) in social recommender systems is an interesting research direction since not only privacy is an essential aspect of social networks but also conventional PPCF techniques of computation-intensive cryptography or data perturbation techniques are not appropriate in real online services. Zhu et al (2014) proposed an algorithm for neighbor based PPCF to protect neighbors and individuals' ratings while Li et.al (2016) presented an algorithm for item based PPCF to protect individual privacy during recommendation.

2.5.5 Graph Analytics & Social Network Analysis

SNA is a term that encompasses descriptive and structure-based analysis, similar to structural analysis (Newman (2003); Kolaczyk(2014)) and it analyzes various characteristics of the pattern of distribution of relational edges and draws inferences about

the network as a whole or about those belonging to it. SNA aims to compare networks, track changes in a network over time, reveal communities and important nodes, and determine the relative position of individuals and clusters within a network (Hansen et.al(2010)). As before stated, mining the content of OSN in conjunction with the network can be useful in efficiently answering sub questions of an analysis such as:

-Do friends post similar content on Facebook?

- Can we understand a user's interests by looking at those of their friends?

In a more detail view, analysis tasks of SNA include the following (Gupta(2016); Ferguson(2016); Lin(2015)):

- Discovering the structure of social network
  - Why and how did it come to have such structure?
- Community Analysis
  - What are the communities in the social network?
- Processes and dynamics:
  - How do information, behavior, and diseases spread?
- Path Analytics
  - What is the shortest path between two nodes e.g. find the best possible route for traffic optimization in smart cities?
- Connectivity Analytics
  - What are the connectivity patterns of edges (e.g. find who are the listeners in a social network)?
- Centrality Analytics
  - What are the important nodes regarding to a specific analysis problem e.g. find who are the influencers colleagues?

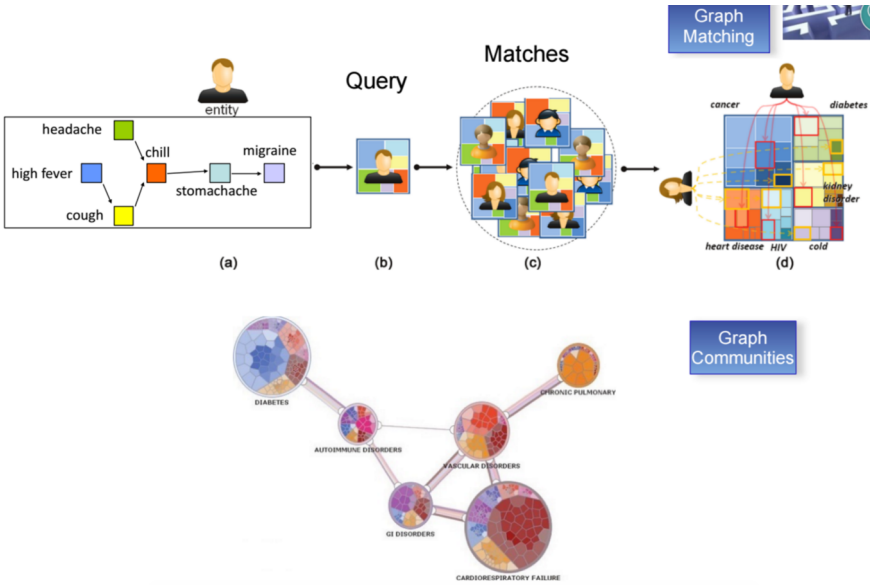


Figure 12: IBM Watson graph matching to allocate symptoms to a disease (Lin(2015))

The following are some applications of social network analysis in Big Data era according to "Big Data Analytics" course offered by Columbia University, supervised by Lin (2015):

- Productivity Growth & Measurement of success
- Finding and Ranking expertise
- Knowledgeable and influential Human Resources appropriate for a project,
- Recommendation
- Customer Behavior Sequence Analytics
- Financial analysis
- Social media monitoring
- Analyzing Trust: Propagating Trust
- Anomaly Detection (Espionage, Sabotage, etc.)
- Fraud Detection
- Cybersecurity
- Web page ranking
- Intelligent computing, as depicted in Figure 12 in IBM Watson.
- Visualization of social roles, as depicted in Figure 13.

Figure 13: Visualization of social roles (Lin(2015))

Graph based mining tools are required in order to easily model the structure of the social networks and perform the above tasks. A comparative study of a range of such tools has already done earlier by Agrawal et.al (2015), Akhtar (2014), Kennedy et.al(2013) Huisman Duijn (2005). In particular Akhtar (2014) compared the following instruments: Networkx, Gephi, Pajek, Igraph in terms of platform, execution time, algorithms complexity, input file format and graph types. One of their conclusions is that stand alone software is very useful for graph visualization, data format conversion and easy to learn; so for beginner Pajek and Gephi are suitable software. For complex dataset and research purpose Networkx and IGraph software are appropriate.

Agrawal et.al (2015) conduct a comparative analysis among five instruments Pajek, Gephi, Netlytic, Social Network Visualizer and Graphviz and included parameters of visualization layouts, graph types, clustering algorithms and dendogram display. They came to the conclusion that all are suitable for chart measurements, degree centrality, closeness centrality and betweenness.

Kennedy et.al (2013) provide brief summaries of different tools range from social network analysis (NodeXL, Gephi) to user generated content analysis (SocialMention) in order to examine their usefulness to public sector organizations for the purposes of public engagement. As such, it does not represent a comprehensive review of the full capabilities of each tool.

2.6 Summary

Most literature conceives data cleaning of unstructured data as a significant area to investigate, given that not only OSN is a rich source of this type of data but almost 80% of the generated data today are unstructured as well. Regarding content based analysis

algorithms such as sentiment analysis and topic detection are subject to scrutiny from a data quality point of view. In general, content based analysis has been extensively studied for product and movie reviews, which differ substantially from online social networking data. Even if recent work analyzing OSNs is growing, the area still presents many open challenges such as the lack of methodologies that adopt a more systemic view that combines approaches both from network effects including nodes and edges formation (mostly graph-structured data) and from content features (mostly unstructured data). Obviously, this demands both a profound theoretical perspective and appropriate tools.

3. METHODOLOGY

The purpose of this chapter is to describe the design decisions and procedures that are chosen to systematically approach the research questions. The conceptual framework being set in the previous chapter, guided the construction of the followed methodology. In section 3.1 a short overview is given about the different types of research together with a short elaboration on steps that construct the methodology design plan. Section 3.2 elaborates on the reliability and validity of this thesis and section 3.3. describes the contribution of this survey.

3.1 RESEARCH TYPES

The paths that lead to knowledge have been shifted, since 21st century science is becoming cyber-science and a new scientific methodology is being born and called "data-driven" (Cukier and V. Mayer-Schoenberger (2013); Laufenberg (2010); Parker and Thomson(2016), Djorgovski(2015), Degryse(2016), Lesk(2016), Hey(2012), Hannay(2015), Markoff(2009)). To gather even more data for this thesis, both survey and experiment was conducted. The main research question of this thesis starts with a what question and an in depth survey is perfectly suited for what questions (Yin (2009)).

Three of the most common purposes of the design research plan are: exploration to discover new connections, description to describe the main aspects of the topic, and correlation to study relationships between two or more variables (Babbie (2007)).
To achieve that, a big amount of literature regarding data analysis in Big Data and Social Networks is surveyed. Social networking data inherent Veracity dimension, led us to the significance of data cleaning. We make an exploratory research in the data cleaning issue in online social networks since the literature shown that this process is not clearly defined in OSN's. Then, we make a descriptive research to provide an accurate portrayal of data analysis techniques related to the features of social network data. We make a correlation analysis between the below variables:

- Data cleaning techniques and error repair
- Data analysis techniques and frameworks' services
- Computing Intelligence techniques and analysis purpose.

In addition to the survey, an experiment was conducted since experiments are perfectly suited for how questions (Yin (2009)). Specifically, two tools that have gained recently a huge popularity are chosen: Google Cloud Natural Language API and Indico Text Analysis API. These tools are employed to conduct sentiment prediction in order to answer two sub questions (i) their accuracy on results and (ii) the relation of the results within formal and informal contexts. These two sub questions help to approach the matter of what is the current state of API tools regarding sentiment analysis in OSN'S. To conduct this experiment, a cross-platform tool with Node.js and MongoDB is built and can be used by anyone who desire to evaluate a textual sentiment through machine learning API's.

Generally, our research is based both on qualitative and quantitative results which are visualized inside comparison tables, figures and charts.

3.2. Research Reliability and Limitations

To receive reliable and valid data and get a better understanding regarding the points mentioned in the methodology section, both the structure of the survey and the experimental environment set up with the tool are conducted with the help of theory.

Regarding to the survey:

- In order to provide a complete view on the research topic, a big amount of literature is surveyed. Each further step is built upon asking the right questions connected to the comprehension of the previous one.
- Every comparison between tools, techniques and frameworks is based on a theoretical perspective and not on a numeric evaluation-benchmarking; because each tool offers a very specific application and even those with a similar objective couldn't be compared in numerical terms since both the absence of benchmarks and norms of standardization and the recent advent of the matter, wouldn't allow us to be objective.

Regarding to the experimental environment:

- First, owing to the API call limitation of our experimental platform, the number of opinions used in system evaluation is limited.

- The lack of public datasets of formal OSN like LinkedIn or Academia.edu force us to compromise with IMDB review datasets.
- Both the lack of evaluation datasets and API call limitations didn't allow us to involve large scale data to our tool although the latter supports them.
- The need for classified text made the search of evaluation datasets difficult. Though, dataset selection does fulfill accuracy dimensions since data are manually classified and used in other scientific papers, as well.

3.3 Research Contribution

There is a need for a theoretical perspective in this area since it is a new research topic and a theoretical research is the trigger and the base for the practical researchers to build tools, algorithms and frameworks. The contribution of this survey is significant for many reasons. First, this survey provides sophisticated categorization of a large number of recent articles according to the data analysis tools and frameworks in OSN's. Big data cleaning frameworks are also searched and studied since the combination of these two stages and not the analysis stage solely, is evidently the one that offer insightful information via data science algorithms and methods. This angle could help the data scientist to choose a variety of frameworks to use for their analysis purpose. We also divide the techniques used in the frameworks and their corresponding limitations if are any; therefore, researchers who are familiar with certain techniques could enhance them for a certain application development or improvement. Additionally, this research is useful for new comer's researchers to the social network and big data analysis field to have a panoramic view on the entire field. Generally, this survey can be useful for everyone who desire to learn about social networking data analysis frameworks, techniques and applications. In addition to the theoretical framework, a tool is developed for discovering the current effectiveness of sentiment analysis in OSN's via API's. The tool is useful for anyone who desires to perform a sentiment analysis through text analysis API's. It is generic, scalable and with high speed allowing for other scientist to analyze sentiments given at least a dataset and an API.

4 Social Networking Data Cleansing

In this chapter is investigated the stage of data cleansing in the context of OSN's. In Section 4.1. an overview of the data value chain and the types of dirty data in social networks are given. Section 4.2. presents the most widely used techniques to handle low quality data. In Section 4.3. is described big data cleaning frameworks and Section 4.4 investigates the contribution of cloud computing in cleansing tasks. Lastly, section 4.5 provides a conclusion.

4.1 Low-Quality Data in OSNs

The analysis of social network data, like any other data analysis, complies with the big data value chain illustrated in Figure 14.

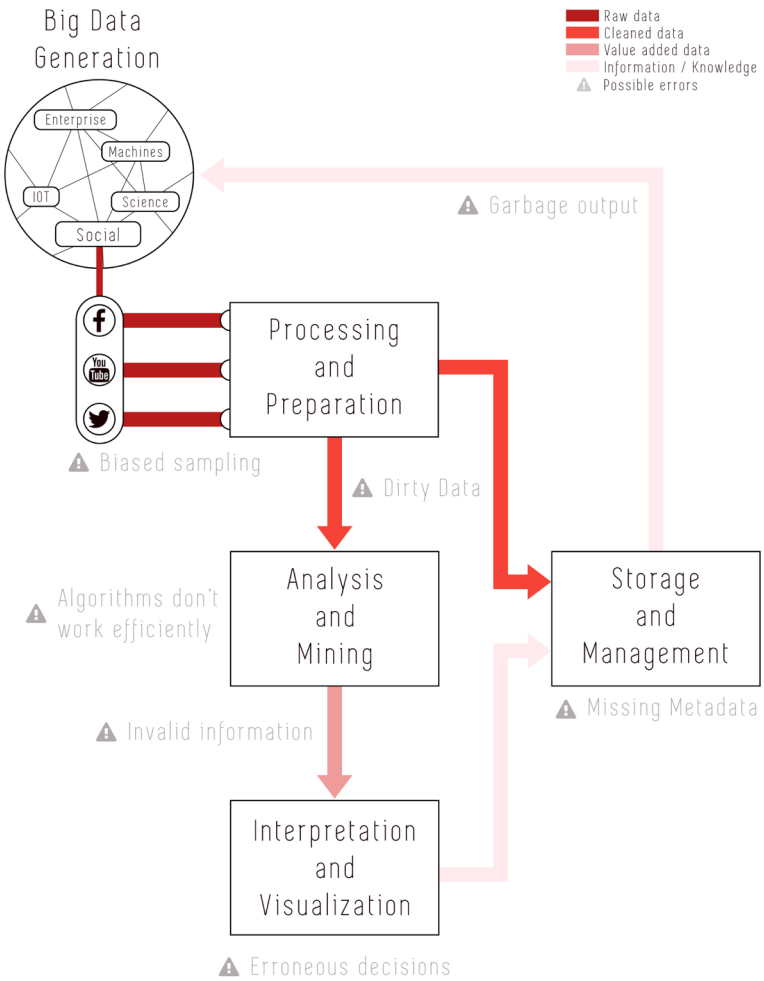


Figure 14: Social Network Data Value Chain

Too much noisy or even faulty input data often lead to a less than the desirable algorithm performance. In addition to that, databases and hardware are susceptible to dirty data (Batrinca & Treleaven (2015)). Therefore, when the quality of data is preserved from the

one stage to the next, the possibilities to discover knowledge increase and so does the value of the data. Possible errors occurred in each stage that hinder the quality, are also illustrated in Figure 14. From a theoretical standpoint, low quality data can occur either at instance level (the data itself) or at schema level (metadata).

4.1.1 Issues at Instance Level

At instance level, dirty data usually is presented in two forms: missing values (MVs) and wrong (noisy) data. Data from online networks may suffer additional problems because of the nodes-edges ambiguity and the prevalence of informal language. All of the studied research in analysis domain, present the strive to extract clean data from OSN. For instance, Sankaranarayanan et.al (2009) acknowledged that the main issue in analyzing tweets to built a topic detection tool, was dealing with noisy data.

Duplicate nodes artificial nodes and inactive nodes are some node-centric issues (Bonchi et.al(2011); Hassan & Menezes(2013)). By inactive nodes, it is described the fact of users who have an inactive account in social media. Edges and other information may be duplicated or missing. Inferring missing attributes of user in OSN were predicted with the usage of Community Detection by Mislove et.al (2010). Incomplete data are caused by imperfect data acquisition process, no authorized access, communication failure or because they are not yet reflected in the online network (Bonchi et.al(2011); Fire et.al(2013); Hansen et.al(2010)). Other irrelevant information that should be omitted are web robots, extensions of CSS, GTF, FLV and the records with failed HTTP request, to name but just a few. Last but not lest, text may contain misspelled words, quotations, extra spaces, extra line breaks, special characters, foreign words and the like which also should be deleted.

4.1.2 Issues at Schema Level

At schema level, social network nodes and edges are ubiquitous, namely different types both in relationships between nodes and in nodes per se are presented (Bhatnagar (2013)). For example, the network of Facebook in addition to friendship relationships between persons, has got relationships of other types, such as person-photo tagging relationships or person-movie liking relationships. The problem is compounded when combining data from multiple social networks; because not only the vastness of diversity in relations becomes even bigger but also a unified conceptual data model that will support the various data structures under a single scheme is missing as well. Furthermore, when

integrating data from multiple social networks it should not be overlooked the fact that, the same user may be presented differently in between the social networks. Each social network site has developed both unique characteristics in text sharing (Hassan & Menezes(2013)) and its own schema-network representation which is full of data siloes (Batrinca & Treleaven (2015)). The differences on vocabulary, some of them depicted in Table 3, arise the need for cross-domain vocabulary matching (Tan et.al (2013)). Besides the above, the responses of APIs are often structurally incompatible between services.

Table 3: Indicators of vocabulary-based differences among three social networks

Social Network	Twitter				Flickr					Facebook			
Nodes / Edges	user	tweet	hashtag	Nodes / Edges	user	tag	comment	image	Nodes / Edges	user	page	photo	post
follow	✔			upload	✔			✔	account	✔	✔		
post	✔	✔		post	✔		✔		album	✔	✔	✔	
reply		✔		comment			✔	✔	like	✔	✔	✔	✔
contain		✔	✔	contain		✔		✔	feed	✔	✔		✔

Understanding the context of data, the node-link types of a network and the underlying limitations are necessary in data cleaning. The need for provenance, privacy and the employment of qualitative data cleaning approaches on distributed streams of data are yet big challenges (Chu et.al (2016)). For instance, how to better preserve the privacy and utility of social network data to benefit data analysis, studied by Wang et.al (2014), is an open issue topic.

4.1.3 Issues arise by extraction process

Although many social media data are accessible through the API, they usually limit the number of API transaction per day. Importantly, not all sites (e.g. Skype, LinkedIn) provide API access for scraping data. Each social platform has very specific rules around how to use their respective data that can be found in the Terms of Service. There are also tools used for scrapping that protect raw data or provide simple analytics such as Google Trends, SocialMention and Social Pointer (Batrinca & Treleaven (2015)).

Scrapping such huge social networks requires robust systems with high processing power and huge storage capacity. Therefore, collecting a large amount of a social network's data is sometimes infeasible and traditionally data analysis is upon a snapshot of OSN. This could lessen the quality of the data if the representativeness of samples to the original full

dataset is inaccurate. In addition to that, each social network includes users that share same interests, race/ethnicity, socioeconomic status and other characteristics which can cause potential biases on sampling (Tan et.al (2013); Chu et.al (2016)). A comparative analysis of data samples representativeness, obtained from social network stream APIs, is conducted by Wang (2014). With regard to sampling linkage data, best practices are proposed by Maiya & Berger-Wolf (2011). Apart from checking both how the "on-line" filters may discard useful information (Jagadish et.al (2014)) and how filters bias the result (Shuguang et.al (2016)), equally important is the constructive communication between the business and technological staff so to acquire only the relevant and the exact amount of data (Parashar and R. Carlson(2015)). Jagadish et.al (2014) also recognized the need for efficient incremental ingestion techniques, since loading of large datasets is often a challenge, especially when combined with on-line filtering and data reduction.

Besides these, many times the imperfect acquiring process provokes missing data to which the common data mining and Machine Learning models are sensitive. Data cleaning comes as a step after the extraction of data to improve the collected dataset. In Table 4, a summary is given about the correlation between the common data quality problems and the errors arise in OSN.

Table 4: Data quality problems in Social networks

Data Quality Problems	Social Networks	
Error Correction	Informal Language: Misspelling, slang, abbreviation	
	Irrelevant Information: Automated programs, spam, extensions	
Conversion	Unstructured data to structured	
	Convert one format (usually XML or JSON) to the one needed for analysis	
Integration	Duplicate or Ambiquity	Vocabulary, network schema, services differ from one to another
		In human language
Missing Values	Links, Attributes	

4.2 Data Cleansing Techniques

Data cleansing is considered as a "black art", a behind the scenes process which often results in undocumented methods (Maletic & Marcus (2009)). Additionally, there is no commonly agreed formal definition of data cleansing problems (Hu et.al (2014)). Arguably, it is a topic that needs a lot research. Methodologies used in data cleaning are

categorized into model-based, namely qualitative techniques, and machine learning or statistical techniques. Note though that, many qualitative techniques are amenable to such statistical analysis and sometimes there is an overlapping; The several methods found in data cleansing process are listed: integrity constrains rules(IC), statistical, pattern-based, neural networks, parsing, association rules and Machine Learning (ML) techniques. Figure 15 illustrates the data cleaning techniques according to three variables: how, where and by whom the data errors will be detected and probably repaired.

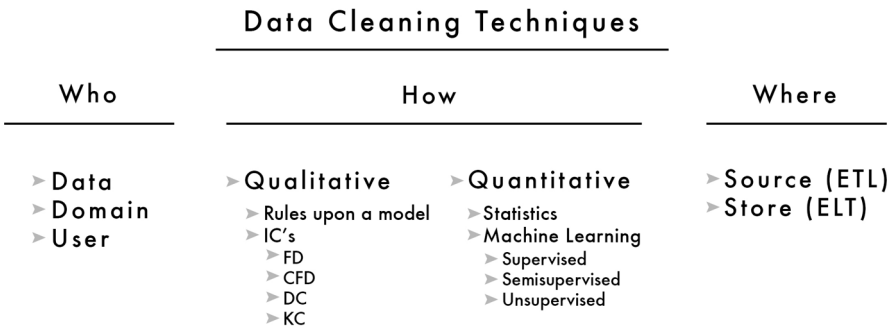


Figure 15: Data Cleaning Techniques

4.2.1 QUALITATIVE PROCESS

Data cleaning mostly consists of two stages: error detection and error repairing (Chen et.al (2016)) but in a more detailed view data cleansing refers to the following process (Khayyat et.al (2015); Chen et.al(2014); Tang(2014); Taleb et.al(2015); Chu et.al(2016))

1. Determine data errors by specifying quality rules;
2. Search and identify error types with regard to the specified rules;
3. Correct errors by updating, repairing or deleting them;
4. Documenting error examples and error types; and
5. Modify data entry procedures to reduce future errors.

In step (1) rules can be specified by the user or crowdsourcing, by domain knowledge such as knowledge bases or by automated machines. With the latter is meant that since there is a vastness of data, the rules need to be learnt from the dirty data itself, centralized or distributed, and validated them incrementally as more data is gathered. However, since data itself is dirty, the need is to make these rules robust against outliers and to allow approximation (Khayyat et.al(2015); Vaidya (2016); Saha & Srivastava(2014)). Each

data quality rule usually targets a specific data quality issue such as duplicate instance rule, illegal value rule and functional dependency rule. But interacting two types of quality rules may produce higher quality (Tang (2014); Chu (2013)).

To define error types, data should be analyzed so as to profile them and mine the rules. Data profiling is used on the instance analysis of individual attributes such as data type and value range and it gathers data structure, data pattern, statistical information and metadata sources for data management. Data mining, a key technique for data cleaning (Natarajan et.al(2010); Maletic &Marcus(2009)), discovers specific intrinsic data patterns. Interestingly, data mining algorithms have been found to be capable of handling the three dominant disputes of noise, size and dynamism of social network data. After data analysis, context-aware rules should be derived from the certain subset of data so as to be applied in step (2), and better still to fix those errors by using the rules.

After step (1), the data cleansing process iteratively runs steps (2) and (3) until obtaining a repair, that satisfies the specified rules. Violation to rules indicate data inconsistency and either one deals with these inconsistencies without repairing them, or finds ways to repair them. Repair also can be performed by using scripts, human crowds, or a hybrid of both. Extremely important in data analysis is the data provenance (Vaidya (2016)) which is accomplished through steps (4) and (5).

Taleb et.al (2015); Saha and Srivastava(2014) Bhatnagar(2013) ; Wang (2014) refer to Big data quality with a focus on conditional functional dependencies as the rules to detect and capture semantic errors. Functional dependencies describe data consistency and several systems take these rules, as input for detecting errors and computing a “clean” version of the data. Regarding accuracy rules, the true values of entities when are absent, a framework is proposed by Cao et.al(2013). Other data quality and cleaning rules are in detail described by Furber and Hepp (2011).

4.2.2 Quantitative Approach

Data is corrected based on statistics over value or on supervised classifiers. Although, contextual models that use statistics and semantics as predictors and learn characteristics of streaming data to detect errors and build a repair, demonstrate favorable outcomes in cleaning, very little research has been carried out (Gill et.al (2015)). Shi et.al (2015) utilize a prediction model based on logistic regression and SVM to clean big data coming

from power grid. In addition to building predictive models, there is a lot of interest in link prediction predicting the true values of missing attributes or links. Fire et.al (2013) predict missing links combining many social networked data via supervised machine learning classifiers trained upon the structural features of the graph topology. KNIME tool runs neural network (MLP), decision tree, and Naïve Bayes classifiers and selects the best threshold value appropriate to fill missing values (Silipo et.al (2014)). In addition to ML, Bag of words and TF are also employed in text mining and network analytics in order to infer customer intelligence from social media (Thiel et.al (2012)).

Aetas system (Abedjan et.al (2015)) uses ML techniques to discover rules that take into account the time dimension of data. Specifically, they extract events from news media and discover approximate functional dependencies with support of temporal dependencies.

Considering the noise dimension generated by automated programs, much recent work focused on spammer detection in social networks via supervised machine learning classification methods. Based on structural features, Tan et.al (2012) detect spammers via Naive Bayes (NB), Logistic Regression (LR), and Decision Tree (DT) whereas Zheng (et.al)(2015) use both structural and content features to train a SVM classifier and identify spam.

Given a wide range of data cleaning practices the first challenging question is which to pick given a specific task (Tang et.al (2014)). As a response to this, Mezzanzanica, et.al (2014) have recently proposed a Universal Cleansing framework to automatically identify the most accurate cleansing technique among alternatives through model-checking techniques and a data-driven policy. Boselli et.al (2015) extend this framework by including ML algorithms trained on the data recognized as consistent reducing therefore the dependence on domain experts. The framework focuses on violation of semantic rules defined over a set of data.

4. 3 Big Data Cleaning

Guidelines for conducting big data processing are the specification of filters in such a way that they do not discard useful information, the automatically generation of the right metadata to describe what data is recorded and how it is recorded and measured, the data

provenance and the distinguish between spam and valid information (Almeida and C. Calistru (2013); Boyd & Crawford (2012)).

Table 5 illustrates the correlation between data cleaning tools and data errors they handle.

Table 5: Services of data cleaning frameworks

Tool	Technique			Data Errors						Data		Service
	Rule-based	Statistical	Other	MV's	Noise	Duplicate	Discovery of Repair	Spam	H.L.I.	U	S	
Fire et.al(2013)		ML		✓						✓		Predict missing links
Gill 2015		Predictive Model		✓							✓	Clean sensor data streams
Shi et.al(2015)		ML		✓	✓					✓		Clean power-grid data
Schmidt et.al (2015)		ML			✓					✓		Preprocess chemical data
Cao et.al	✓				✓						✓	Determine accurate value
Knime (2012),(2014)		ML		✓	✓	✓			✓	✓	✓	Analytics & Preprocessing Platform
Chu et.al (2015)			Knowledge Base		✓						✓	Data cleaning system
Hassan & Menezes (2013)			Graph random walks		✓				✓	✓		Social media text normalization
Nirmal & Amalarethinan (2015)			✓		✓				✓	✓		Preprocessing framework for sentiment analysis
Mislove et.al			Community Detection	✓						✓		Inferring user's attributes in OSN's
Peled et.al 2014		ML, ER				✓				✓		Matching entities across OSN's
Mezanzanica et.al (2014)	✓						✓					Universal Cleaner, a system finding cleaning solution
Boselli et.al (2015)	✓	ML					✓					Enhance Universal Cleaner
Volkovs et.al (2014)	✓	Logistic Classifier					✓				✓	Adaptability when constrains changed
Abedjanet et.al (2015)	✓						✓			✓		Discovery of events' duration
Ebaid et.al 2013	✓						✓			✓		Nadeef: data cleaning system
Taleb et.al (2015)	✓						✓			✓		Preprocessing framework for data quality
Immonen et.al	✓						✓			✓		Architecture for quality management for OSN
Tan et.al (2012)		ML						✓				Detect spam posts in OSN
Zheng(et.al)(2015)		ML						✓				Detect spam posts in OSN

MV'S: Missing Values
H.L.I.: Human Language Informality
U: Unstructured
S: Structured

The usage of Entity Resolution (ER) can also effectively improve the quality of big data sets and reduce the number of samples hence enhancing the speed and quality of data analysis (Tsai et.al (2015)). ER, also called entity matching or domain knowledge, usually compare pairs of entities by evaluating multiple similarity measures and can be either supervised or unsupervised. The former needs labeled training datasets or predefined thresholds to base their decisions on whereas the latter avoids human intervention by using clustering algorithms that group together items which present a high similarity. ER can be further divided into semantic-based and syntactic-based similarity approaches. The first measures how two values, lexicographically different, are semantically similar whereas the second computes the distance between two values that have a limited number of different characters. Google Refine, recently called Open Refine due to open source

availability, has leveraged Freebase to perform Entity Resolution. Recently though Freebase was replaced with Knowledge Graph API (Starr (2016)).

Schmidt et.al (2016) provided an overview of methods for preprocessing structured and unstructured data in the scope of Big Data. Specifically, they used NLP techniques such as the representation of TF-IDF implementation of the-bag-of-words model, POS, Machine Learning techniques and the filtering of stop words. Spark Streaming, an extension of Apache Spark core API, used to clean, analyze and visualize social media data in real time employing K-means algorithm to cluster tweets. It is worthwhile to mention that, anonymization of the data at an early step should be considered and when automation is the case then the possible loss of information should also be tackled (Schmidt et.al (2016)).

4.3.1 Cleansing Unstructured Data

Sophisticated techniques for the preprocessing of unstructured data have been proposed in the research areas of Information Retrieval, Machine Translation and NLP and have been widely adopted in the Text Mining community. The process of cleaning is heavily depending on the analysis purpose and therefore not all scientists follow the same procedures in discarding unimportant data.

Li and Li (2013) utilized a POS tagger, Mithun (2012) employed rule-based patterns and regular expressions and Prom-on et. al (2016) used tokenization and removal of outliers. In a trend analysis system, Sociopedia (Ramachandran(2015)), stop words tweets with too many hashtags (#) and @ (considered as spam) are eliminated in filtering step. Zhou et.al (2015) and Sankaranarayanan(2009) create a keyword lexicon to filter out irrelevant to their analysis tweets. Zhou et.al (2015) pre-processed tweets by time expression resolution, NER, POS tagging and stemming, and finally the mapping of named entities to semantic concepts whereas Sankaranarayanan(2009) employ TF-IDF.

The problem of users to use a different slang or acronyms for almost every common word in English can be addressed using NER and vector space model. Zhou et.al (2015) leverages Freebase to enable NER and categorize events into semantic classes. Additionaly, Peled et.al (2014) use NER to administer the problem of identifying different profiles, which belong to the same individual between social networks.

Note that, POS and NER methods use sentence structure and language features whereas TF-IDF, measures the significance of words from text ignoring sentence structure. The goal of POS, tagging data with metadata and other preprocessing techniques is to give unstructured data a structure, to create patterns and/or reduce ambiguity for subsequent language analysis. This makes content easier to be searched, to be analyzed and to be shared (Batrinca & Treleaven(2015); Schmidt et.al(2016)). Another recent work regarding sentiment analysis is the proposal of a framework for effective pre-processing of Twitter Feeds by Nirmal and Amalarethinam (2015). In particular, researchers employ text mining to build a clean text corpus for normalization of data.

To efficiently handle typographical errors character-based similarity, token-based and phonetic similarity metrics can be employed (Peled et.al(2014)). A social media text normalization system that corrects noisy words and typographical errors is presented by Hassan and Menezes (2013). They propose an unsupervised approach that learns the normalization candidates from unlabeled text data and maps the noisy form of the word to a normalized form. Their technique utilizes a normalization lexicon based on distributional similarity (semantic ER) and string similarity (syntactic ER) via random walks and it is adaptable to any domain and language.

4.3.2 BIG DATA CLEANSING FRAMEWORKS

Dedoop (Kolb et.al(2012)) is an entity matching framework based on MapReduce and cloud for the purpose of parallel deduplication of large datasets. Chu et.al (2016) also develop a distributed big data cleaning system for data deduplication. Another state-of-the-art framework, *Katara (Chu et.al (2015))*, bridges crowdsourcing and Knowledge Bases to find table patterns that can align information at the instance level to achieve reliable data cleaning. *SampleClean* (Wang et.al (2014)) employs data cleaning to mitigate errors of sampling in query results upon large datasets. *BigDansing (Khayyat et.al (2015))* takes integrity constrains (IC) into a series of transformations that enable distributed computations and several optimizations with focus on scalability. *NADEEF* (Abaid (2013)) is a cleaning system that leverages the the separable execution of two main tasks: (1) isolating rule specification; and (2) developing a core that holistically applies these routines to handle the detection and cleaning of data errors. Another data cleaning framework is *LLUNATIC (Geerts et.al (2013))* which develops parallel-chase procedure that chase violations of rules and guarantees both generality and scalability. Both NADEEF and LLUNATIC are effective for static data and fixed constraints, and to

overcome these constrains, Volkovs et.al (2014) propose a cleaning framework for dynamic environments. They presented a classifier that predicts the type of repair needed to resolve an inconsistency, and automatically learns user repair preferences over time.

4.4 Cloud-based Data Cleaning

Although Big Data analysis can also be deployed in non-cloud clusters of computers, cloud represent the "natural" context for them because it performs massive scale and complex computing in a cost-effective manner. However, regarding preprocessing the major problem would be getting the data into the cloud to begin processing (Ahuja and Moore (2013)). To this extent, to take advantage of cloud for big data analysis, data must in the first place be in cloud. To address the issue of uploading big data to the cloud, a number of approaches to WAN optimization have been established. The techniques include: compression, data deduplication, caching, and protocol optimization (Raj and Pethuru (2014)).

Many solutions need to be found with regard to data management and data processing in the cloud. Some critical challenges include the roles of humans on data life cycle like how to support essential services such as data curation and provenance and how to identify relevant information sources and incrementally refine the data processing pipeline (Abadi et.al (2016); Hashem(2014)). Data are often collected from different sources which provokes the serious problem of poor quality data for many cloud service providers (Tan et.al(2013)); Hashem(2014)). Being unable to evaluate every data item on its validity given the volume is a huge obstacle. Solutions to these problems could be processing data in the source or identify quality data such that only a subset is required to be retrieved (Ahuja and Moore (2013)). High-quality data in the cloud is characterized by data consistency; namely if data from new sources are consistent with data from other sources, then the new data are of high quality (Hashem (2014)).

Cloud computing technology has been applied in the field of ML but there is still no real application to Big Data cleaning algorithm (Hamami et.al (2015). For instance, in a case study (Jagadish et.al (2014)) the data dumped into a cloud platform, after the cleaning phase has been performed, for further analysis which made use of ML algorithms. Google Cloud Natural Language API (Google Cloud Platform) reveals the structure and meaning

of text by offering powerful machine learning models in an easy to use REST API. It is used in data preprocessing with syntax analysis like POS tagging and Entity Recognition. However, these algorithms are learnt to work with conventional data sets such as news media and web pages restricting its use on Social Networks. Map-reduce as a distributed processing model isn't suitable for iterative processes and data cleaning is a highly iterative process; Spark, however, deals with iterative processes effectively but with a clear impact on speed (Reyes-Ortiz et.al (2015)).

4.5 SUMMARY

To summarize, there is no universally applicable data cleaning method and when selecting algorithms for a given dirty data set, several basic factors have to be considered such as the nature of outliers, robustness, the existence or not of a clean and complete data set that can be used as a training data set and the efficiency vs accuracy trade-off. Many social networking sites have between 10 and 200 million users, so data sampling is central to most studies. As a hint to the question which portion of data is relevant, one should keep in mind the context and its limitations. Among the studied frameworks only a few cope with the informality and 'dirtiness' of human language.

Data cleaning in social-networking data analysis is even more laborious comparing to other unstructured data sources and normalizing unstructured data is still a highly computational intensive and time-consuming task. Machine learning algorithms, widely used for the cleaning task, have high computational cost and therefore have recently applied in cloud computing. To the best of our knowledge, there are many cloud-based tools for transformation, integration and analysis in the market but not for cleansing.

Although human evaluators have been surpassed by machines in many fields, the guarantee of the accuracy of data cleaning process without verifying it via experts or external sources is yet infeasible. In order to achieve more accurate analysis results questions like: how the determination of data trustworthiness, how the identification of errors and how the biases are evaluated and corrected are important to be asked; and when are answered to be recorded so as to keep data provenance and spot possible mistakes either human or machine generated.

5 Social Network Analysis (SNA)

The purpose of this chapter is to relate graph analysis metrics with four SNA tools. In section 5.1 centrality analytics that define users' influence in OSN are described while section 5.2 presents link analysis algorithms and applications. A brief introduction about path analytics and community detection is given in sections 5.3 and 5.4 respectively. A comparative analysis between data analysis and platform requirements among NodeXL, Networkit, Pajek, Gephi and Statnet is provided in section 5.5. Last the conclusions are illustrated in section 5.6.

5.1 Influence Analysis

In the graph community, centrality metrics are typically used for measuring the dominance of such nodes, quantifying the strength of connections and uncovering the patterns of influence diffusion. The relatedness of centrality measures to OSN is illustrated in Table 6 and Figure 17. In OSN a critical research topic is to identify 'experienced' or 'trusted' users that may be trendsetters since their opinionated posts are the one that can rapidly spread far and wide in the network enabling them to influence other users. An interesting fact regarding trendsetting is that, how much credence another person gives a post may depend on how many times they hear it form different sources (flow) and not how soon they hear it (geodesic distance) (Hanneman and Riddle (2005)). Identification of influential users and of whether individuals would still propagate information in the absence of social signals about that information are two elements required to be studied in order to study information flow in OSNs (Bakshy et.al (2013)).

Table 6: Centrality Interpretations

Interpretation of Centrality metric in OSN	
Degree	How many people can this person reach directly?
Closeness	How fast can this person reach everyone in the network?
Betweeness	How likely is this person to be the most direct route between two people in the network?
Eigenvector	How well is this person connected to other well-connected people?

In opinion mining framework, proposed by Prom-on et.al(2015) Li and Li (2013), is utilized the degree centrality to determine influential users in Twitter microblogging service, as shown in Figure 16. In addition to centrality metrics the evolution of

topological measures in the network is important indicator of trustiness. Moreover, influence analysis has been considered in recommender systems since friends have a tendency to select the same items and give similar ratings (He and Chu (2014)). In this degree, social recommendations could be move to broader models based on trusted-friends communities of interest, except for the specific user models based on individual behavior and personal tastes (Agreste et.al (2015)). However, noted that, different definitions have been given to what an influential user is.

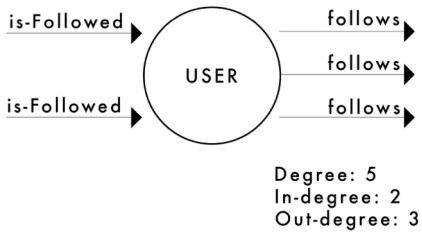


Figure 16: Simulation of an influencer node in Twitter's network directed graph

Agreste et.al (2015) in investigating trustiness in OSN explained the centrality measures well. *Closeness centrality* CC_u measure require to consider the distance between two vertices *u* and v, defined as the length SP (u, v) of the shortest path (geodesic distance) connecting them. It is defined as the reciprocal of sum of all distances from v to all other vertices in the network:

$$CC_u = \frac{1}{\sum u \in V^{SP(u,v)}}$$

Given any three distinct vertices v, u and w, let σ_{uw} be the number of shortest paths from u to w and let $\sigma_{uw}(v)$ be the number of the shortest paths from u to w passing through v. The *Betweenness Centrality* BC_v of v is defined as follows:

$$BC_v = \sum_{v \neq u \neq w \in V} \frac{\sigma_{uw}(v)}{\sigma_{uw}}$$

Other centrality indices are based on the computation of the eigenvectors (and eigenvalues) of the matrix representation of G. The first case is the *Eigenvector Centrality EC*, which defined by means of the adjacency matrix A for G. Let $\vec{x}$ be a |V |-dimensional column vector that satisfies the following equation:

$$A\vec{x} = \lambda\vec{x}$$

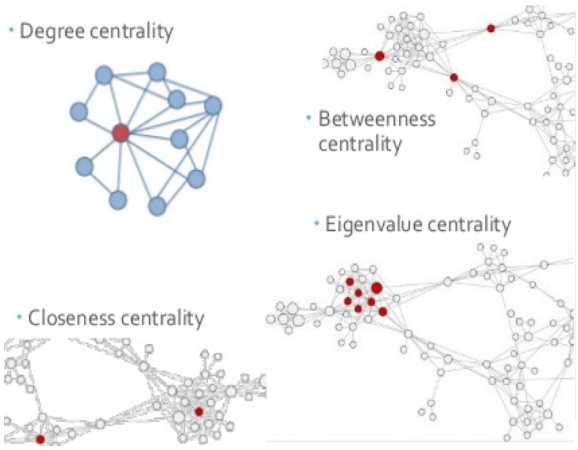


Figure 17: Centrality Measures in Networks

5.2 Link Mining

Link-analysis is used to evaluate connections between nodes. Understanding the formation and evolution of such connections in social networks requires longitudinal data on both social interactions and shared affiliations (Kossinets and D. J. Watts (2006)). Link mining is usually associated with text mining and can be used for classification, prediction, clustering or association-rules discovery. It is applicable in collaborative recommendation systems to identify a group of friends with similar interests. PageRank is the famous link-analysis algorithm used by Google to order search engine results. However, recently Google (Metz (2016)) announced the replacement of PageRank with a more efficient search algorithm BrainRank which is based on Deep Learning Networks. PageRank and HITS algorithms are also used in influence analysis. Both of the two were used in a sentiment-analysis framework by Li and Li (2013) to evaluate the credibility of an opinion-expresser on Twitter. Khalid et.al (2014) utilize HITS to rank the most experienced users in venue recommendation. Chen et.al (2016) have evaluated the

performance of PageRank in personalized recommendation whose performance was substandard.

Note that, geodesic distance has been found to be one of the most significant features in link prediction (Fire et.al (2013)).
Geodesic distance and path analysis is used to identify all the connections between a pair of entities, useful in understanding risks and exposure of a network (Gupta et.al (2016)).
Power law, a statistical metric like the 80-20 rule, used to check if a network follows a scale-free distribution of connections to nodes and it is useful for defining popularity (David and Jon (2010)). In a scale free network there are a very few nodes ('hubs') that have connections much bigger than the average degree; while the network grows, these nodes will continue to get a larger share of new connections (Bandyopadhyay et.al (2011)). Twitter and Facebook are distinct examples of such network structure (Li and Li (2013)).
Connected components is another interesting statistical metric that allows the study of dissemination of information in a social network. A connected component in a graph is referred as a set of nodes and edges where there exists a path between any two nodes in the set (Aggarwal (2011)).

5.3 Path analytics

Path analytics usually approach optimization problems like finding the best possible path (dependencies) between nodes (variables) in a network (set of variables). It is widely applied in business intelligence as part of behavioral analytics. To illustrated this with an example, Google Analytics use path analysis functions to determine how many visitors reach a certain destination page.
Graphical models are powerful tools that can be used to model and estimate complex statistical dependencies among variables. A popular example is the Bayesian network used by Zhou et.al (2015) to explore events shared in the social network of Twitter.

5.4 Community Detection

Communities constitute an important aspect of networks and they are important for both exploring a network and predicting connections that are not yet observed (Alhajj and Rokne (2014)).
Community detection or community analytics is essentially a data clustering problem,

where the goal is to assign each node to a community or cluster in terms of an interaction pattern. The analysis can be categorized in terms of the time dimension (Gupta et.al (2015)), namely:

- static analysis: 'what are the communities at time T?',
- temporal analysis: 'how did this community form?',
- or predictive analysis: 'how a community will grow?'.

One way to define a community is by structure, e.g. communities as cliques. Clique or complete graph is a graph where every node is connected to every other node in the clique. Another pattern in relationships is to discover the degree to which an actor exists in a tightly bound group or if they have connections outside their own group. To explore such a notion of network clustering, dyad and triad census have been utilized (McCranie (2015)). A dyad is a sub graph that represents a pair of actors and the possible edges between them whereas a triad consists of three nodes and the possible edges among them.

Wu et.al (2013) have found as the most important measures for detecting communities to be Degree, Betweenness centrality and Clustering coefficient. The latter assesses the tendency of vertices in a graph to form close-knit groups (Agreste et.al (2015)) and it is defined as the ratio of the number of closed triplets in G graph to the total number of triplets of G. To illustrate this, any three vertices u, v and w form a triangle when (u, v), (v, w) and (w, u) are in *E* set of edges; when there are at least two edges among the vertices, they form a closed triplet.

Graph density is defined as the ratio of the actual number of edges to the number of all possible edges.

Lead-follower algorithm (Shah and T. Zaman et.al(2010)) is a community detection algorithm based upon identifying the natural internal structure of the expected communities. It is used by Vakali et.al (2013) for clustering tweets with the same content. Community detection has also been used to infer information about users from OSN given a set of "seed" users (Mislove et.al (2010)).

5.5 SNA Tools comparative analysis

Graph databases such as Neo4j, graph analytics such as semantics, graphical models such as deep learning and graph mining tools such as Networkit are being developed in order to efficiently handle the need of knowledge extraction from networks.

There is also a great variety of software tools to analyze properties of nodes and edges in a network. Some of the tools were originally developed for network visualization, and now contain analysis procedures and other were specifically developed to integrate network analysis and visualization. Though, a tight integration of social network statistics and visualization is necessary for effective exploration of social networks (Aggarwal (2011)). Each tool has certain strengths and limitations thus opting the appropriate one for a particular task is still a challenge. A comparative study of social network analysis tools has already done earlier by Agrawal et.al (2015), Kennedy et.al (2013), Huisman Duijn (2005), Akhtar (2013) but not in a data-centric approach. Thus, we have added comparative results concentrating on data analysis features taking into account recent advancements of tools, as shown in Tables 7 and 8. Both commercial and freely available packages are considered; business or academic oriented tools are examined, as well. Software application with GUI packages (e.g. Pajek) are easier to learn, while packages built for scripting/programming languages (e.g. Networkit) are more intricate, powerful and extensible. Table 7 presents the comparison of the four network analysis tools based on platform characteristics and the most primary analysis needs in response to user's skills. Table 8 presents a comparison of analytical capabilities according to criteria mentioned in sections 5.1, 5.2, 5.3 and 5.4. We opt for studying metrics and algorithms that utilized in prevalent OSN analysis methods. In Table 8 the different algorithms are differently colored depending on which analysis method they belong. Noted that when a cell contains two values such as “M-L” means that the tools provides the concerned metric in a scale from medium to low.

NodeXL(https://nodexl.codeplex.com/) is a free, open-source template for Microsoft Excel that simplifies basic network analysis and visualization tasks and supports analysis of social media networks for noncoding user *(Smith (2013))*. It is similar to Pajek and Gephi but differs in its ability to directly harvest data from social networks (Kennedy et.al (2013)). Though, Gephi is more flexible in terms of visualization. However, network metrics computation in NodeXL can be slow, so research efforts on improved algorithms, parallelization of execution using multiple processors, and the use of specialized graphic co-processors to speed computation are important. Their future plans include cloud computing techniques in order to compute network clusters efficiently. Improved centrality metrics for directed or bipartite graphs and graphs with varying edge weights are also needed (Smith et.al (2009)). NodeXL supports sentiment analysis of textual data by measuring frequency of words occurrence (Hai-Jew (2015)).

Table 7: Comparison of SNA Tools

Program	**Pajek** Akhtar(2014); Mrvar & Batagelj (2016) Wambeke et.al (2014)	**Gephi** Heymann & Grand(2013); Akhtar(2014) Cherven (2013)	**NodeXL** *Smith (2013);* Hai-Jew (2015) Smith et.al (2009); Hansen et.al(2010)	**NetworKit** Staudt et.al(2014); Kurka et.al (2015)	**Statnet** Kolaczyk (2014); Handcock et.al.(2008); Butts (2016)
Platform	Windows	Windows, MacOS, Linux	Windows Excel	All	All
License	Free* *for no-commercial use	CDDL GNU Free	Microsoft, Free, *commercial version available, http://www.smrfoundation.org/nodexl/	MIT	GPL
Version	4.09	0.9.1	332	4.0.1	2016.9
Package	GUI	GUI	GUI	Scripting language Python	R
Extensible	L	H	M	H	H
Expectable Computing Time	M	M	H	L	M
Objective	*"The network calculator, large data exploration"*	*"An interactive visualization tool; like Photoshop but for graph data"*	*"Simple Network Analysis for social media"*	*"A high performance large scale Network Analysis"*	*"An integrated set of tools for the visualization, analysis, and simulation of network data"*
Easy to use	M-L	M-H	M-H	L	L
Quality Graphics	L	H	M	L	M
Analysis Capabilities	H	L	M	H	H
Large Network	H	L	L-M	H	H
Orientation	Business Academic	Academic	Business	Academic	Academic
Support	Books, Manuals, Articles	Online, Book	Online, Books, Manuals, Articles	Online	Online, Manuals, Articles

L: Low
M: Medium
H: High

NetworKit (https://networkit.iti.kit.edu/), a Python module, is a generic toolkit for high-performance network analysis with efficient graph algorithms, many of them allow parallel execution to quickly process large-scale networks. Its aim is to provide tools for the analysis of large networks in the size range from thousands to billions of edges and

intend to be much faster than the mainstream alternatives (Staudt et.al(2014)). Usability and integration with Python libraries for working interactively with data is also provided. It is a tool comparable to *NetworkX* and *igraph* Python packages which are examined by Agrawal et.al (2015) and Akhtar (2013) albeit with a focus on massive networks, faster execution of algorithms, parallelism and scalability. Though Networkit functionalities are not as comprehensive as NetworkX and igraph according to Kurka et.al (2015). Pajek offers similar data analysis capabilities and network visualization features to NetworKit Staudt et.al(2014)*)*.

Pajek (http://mrvar.fdv.uni-lj.si/pajek/) is a general graph analysis tool for analysis and visualization of large networks. It provides an excellent range of metrics beyond social network analysis routines like various partitioning schemes, cliques, clusters, components and many other features *(*Mrvar and Batagelj (2016)*)*. This tool has been in the market for 20 years and has enhanced its features justifying the extensive use both in academic research and in well-known companies such as *Deutsche Bundesbank and Volkswagen*. However, it only runs on Windows platform and it is relatively weak on visualization. *Pajek-XXL*(http://mrvar.fdv.uni-lj.si/pajek/pajekman.pdf) is a special edition of Pajek for analysis of huge networks.

Statnet (https://cran.r-project.org/web/packages/statnet/index.html) is a suite of software packages like *ergm* and *network* for statistical network analysis in R programming language that implements recent advances in the statistical modeling of random networks. It depends on the set of these core packages to provide its basic functionality for static and dynamic network modeling and is used from the R command line or the recent GUI for less experienced users. What differs between statnet and the other tools is that its focus is on statistical modeling of network data. It is utilized for model estimation, model evaluation and model-based network simulation such as latent space and latent cluster models. All of the models are powered by a central Markov chain Monte Carlo algorithm that can easily handle networks of several thousand nodes or more.

Gephi(https://gephi.org/) is a standalone software that studies the correlation of node properties and network structure by using visual patterns and it supports classic data mining algorithms of Social Network Analysis (Heymann and Grand (2013)). Gephi allows very easy graphical representation of the 'connectedness' (degree), 'influence' (betweenness centrality) and community membership of individuals within a network.

Table 8: Comparison of SNA Tools' Analytic Capabilities

Descriptive analysis- Centrality Analysis- Link Analysis- Content Analysis

Program	**Pajek** Akhtar(2014); Mrvar & Batagelj(2016) Wambeke et.al (2014)	**Gephi** Heymann & Grand(2013); Akhtar(2014); Cherven (2013)	**NodeXL** *Smith (2013);* Hai-Jew (2015) Smith et.al (2009);	**NetworKit** Staudt et.al(2014); Kurka et.al (2015)	**Statnet** Kolaczyk (2014); Handcock et.al.(2008); Butts (2016)
Density	YES	YES	YES	YES	YES
Clique	YES	YES	YES	YES	YES
Flow	YES	NO	NO	YES	YES
Network Diameter	YES	YES	YES	YES	YES
Geodesic distance	YES	YES	YES?	YES	YES
Census	Triad	Triad Dyad	Triad Dyad	Triad Dyad	Triad Dyad
Power Law	YES	YES	YES	YES	YES
Connected Components	-	YES	YES	YES	YES
Degree	YES	YES	YES	YES	YES
Betweenness	YES	YES	YES	YES	YES
Closeness	YES	YES	YES	YES	YES
Eigenvector	YES	YES	YES	YES	YES
Clustering Coefficient	YES	YES	YES	YES	YES
PageRank	NO	YES	YES	YES	YES
HITS	YES	YES	YES	YES	NO
Community Detection	YES	YES	YES	YES	YES
Text mining	-	Plugin Alchemy API	Sentiment Analysis	Python Libraries (e.g. TextBlob)	R Packages (e.g.tm)

5.6 SUMMARY

There are plenty of graph analysis tools, each with their own features and benefits. However, social media network data collection, scrubbing, analysis, and display tasks still requires a remarkable collection of tools and skills. In addition to that, in the case that one needs particular investigation, programming or correlative code improvements may be required.

From table 7, it is referred that Networkit Pajek and Statnet can be used for more sophisticated analysis and between the two easier to learn is Pajek but more updated is Networkit. On the contrary, Gephi can be used when attractive and powerful

visualizations of the network is needed. Last but not least, NodeXL can be used for social media analysis supporting the standard analytic and visualization features. According to the comparative analysis among their analytical capabilities as showed in table 8, centrality and descriptive analysis and the basic algorithms of link mining are supported by all tools; while only recently tools' functionalities are enhanced with content analysis. This is quite reasonable since these tools are used for graph manipulation and statistical analysis of networks until now without combining multimedia content analysis.

To summarize, network approach is one way to analyze data from social networks in terms of network structure, community detection, influencer nodes and optimized paths. Although social network theory is not a new topic of research, recently there is a lot of interest in order to fully explore analysis of OSN's and applications as are listed:

- find users with similar interests or trusted friend's communities in collaborative recommendation,
- detect experienced users in sentiment analysis and recommendation,
- reveal the credibility of an opinion-expresser in opinion mining,
- cluster tweets with the same content,
- detect how tweets are shared and generally how information is disseminating,
- infer information about users in the social network.

6 TOPIC DETECTION & TRACKING (TDT)

This chapter presents an analysis of TDT frameworks that used to derive trends, topics, news and events from OSN. Analysis features range from trend detection service to techniques utilized to build the given tools. All features are based in the theoretical baseline defined in sections 2.5.1 & 2.5.2. The results among the several tools are illustrated in Tables 9 and 10.

6.1 TDT FRAMEWORKS

Trends are typically driven by emerging events, breaking news and general topics that attract the attention of a large fraction of social media users. Currently a large number of social media analytics tools focus on detecting emerging topics. Besides the differences among the tools, it is difficult to compare them due to there is no widely accepted benchmark or measure for the quality of trend detection (Atefeh & Khreich(2015); Mathioudakis and Koudas(2010)). Though a comparative analysis is not in the scope of this research, we thought it significant to study the techniques used in trend analysis frameworks in the context of OSN. A summary of the methods for TDT analysis frameworks is depicted in Table 9 and 10.

Tables 9 and 10 present a categorization of these tools according to the year they were created, the type of detection service they offer, the detection approaches and techniques they employ, whether the tools support real-time applications and other less substantial elements of interest, all of which are described below:

1. "Year" refers to the year the tool was created.
2. "Trend Detection Service" demonstrates the service provided by the tool.
3. "Approach" is based on the theory described in 2.5.2 and indicates whether a tool follows the feature based approach where TF-IDF method is usually used; or the document based where a lexicon resource is utilized.
4. "Techniques" shows the specific analysis techniques used to develop the tool.
5. "Real-time" refers to whether the tool tackles the challenge of real-time topic detection. In trend analysis this is a strongly desired requirement.
6. "U.T.D." and "S.T.D." stands for unsupervised and supervised topic detection respectively and they are inspired by the categorization done in Panagiotou et.al (2016) & Atefeh and Kreich (2015). The assignment of each tool in "U.T.D." and "S.T.D."

indicates whether the detection process, involving clustering and noise separation, occurred in a supervised way (labelled data), in an unsupervised way or in a hybrid way combining both of the two.

7. The field "Additional Features" refers either to user experience or extra analysis services provided by the tool.

8. "Similar to" points out other tools that they are similar to the concerned tool. This information was extracted either by the creators of the tool or by researchers that described the tool.

9. "OSN" denotes the OSN each tool built for and tested on.

10. Last but not least, the "contribution field" is determined through the contribution that each paper claim to make with developing the corresponding tool. TweCom is the only tool presented here that can be used by analysts, after the trend is detected, for further analysis. "OSN" denotes the OSN each tool built for and tested on.

TweCom (Cagliero and Fiori (2013)) is a data mining framework for investigating the most relevant trends in terms of content propagation. It extracts linked tweets with an ad-hoc crawler and provides relations/rules about both content and context. To generate taxonomies from both post content and contextual features (temporal and spatial) hierarchical clustering and aggregation functions were used. For each cluster the keyword characterized by the highest TF-IDF value. The tool extracts the relationships between tweets through generalized association rule mining. The latter is used when general semantics are required. An association rule is an implication $X \rightarrow Y$, where X and Y are item sets, whereas in generalized association rule A and B are disjoint generalized item sets, namely having no attributes in common. The extraction of generalized association rules is performed by means of a two-step process: (i) frequent generalized item set extraction through Genio algorithm and (ii) rule generation from the extracted frequent item sets through the RuleGen algorithm. The latter belongs to CART algorithms and determines statistical relationships between many data layers in order to produce a binary decision tree. Ranking and selecting the most valuable rules is constrained by either (i) the rule schema (i.e., the attributes that have to appear in the rule body or head), or (i) some specific rule items of interest. Analysts can then apply drill-down or roll-up queries to study the temporal evolution and geographical distribution of specific terms. Note that, hierarchical clustering produces a set of nested clusters organized as a tree, called dendrogram, over data and in this case it is employed to discover hierarchical relationships among keywords. Researchers utilize the agglomerative approach where

each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

TwitterMonitor (Mathioudakis & Koudas (2010)) is one of the earliest works in the field of detecting emerging topics on Twitter. Researchers propose both bursting and clustering algorithms which are implemented in the core application. Trend analysis is conducted by identifying bursty keywords (seeks bursts in the popularity of single keywords) or keywords that are often encountered in the same tweets with the bursty ones and group them into trends with keyword co-occurrences based clustering. Specifically, given the grouped keywords into disjoint subsets {Kti}, a trend is identified by a single subset Kti, where Kt represents a set of bursty keywords computed at every moment t and k ∈ Kt and Kti is the subset of Kt. Regarding clustering, a few minutes' history of tweets is retrieved for each bursty keyword, and keywords that are found to co-occur in a relatively large number of recent tweets are placed in the same group. The system applies contextual features of tweets for providing an accurate description for each trend and an interactive UI, where a user can rank and submit their own description, is also available.

Another interesting system for keyword-based event detection is presented by Nguyen and Jung (2016). Apart from keyword frequencies, it takes into account both the speed and number of participants that the propagation of tweets follows**.** They extend tf-idf to score term importance with a corpus of messages which is considered as a document rather than a tweet alone. Given a corpus that belongs in the *i*th sample, collected at a window of time T_k , the keyword score for a certain term w is defined as:

$$S_w(k,i) = g\left(S_{w,c_1}(k,i), S_{w,c_2}(k,i), \dots, S_{w,c_j}(k,i), \dots\right) \forall_{c_j} \in C$$

where S_{w,c_j} is the particular score of the term w that is considered for the context feature c_j.

Researchers consider three context features the degree to which keywords appear over a given time; the diffusion-degree, and the diffusion-speed at which the information spreads from a user to followers. Instead of using lexicon-based or machine learning techniques, they build a semantic network whose nodes are tweets that include these meta-information and the edges between tweets infer their closeness relationships with cross-correlation function. They apply density-based spatial clustering to the semantic network of tweets in order to determine the potential clusters. To cluster similar tweets into clusters, they relate them in terms of time and keyword occurrence frequencies

between the groups of tweets.

Cloud4trends (Vakali et.al (2012)) also detects trends via exploiting keyword frequency TF-IDF and specifically by assigning more weight to terms presented in titles and tags of posts. Clustering is similar to that employed in TwitterStand. Though, instead of applying a fixed-threshold based method that sets as inactive clusters after a predefined period of time, such as in TwitterStand, it dynamically observes the clusters' updating rate and can identify trends at their peak and detect the topics that are no more trending. Also TwitterStand examines the geographical scope of the resulting clusters as a post-analysis process whereas cloud4Trends separately collects and clusters tweets that pertain to a desired geographical area and takes into account the respective user's physical location. The concurrently collection and processing of streams for the different geographic areas offers real fast analysis. In particular, it collects data from three different sources namely tweets, blogs and extended tweets and processes them in the cloud using the MapReduce paradigm.

*TwitterStand (*Sankaranarayanan et.al (2009)*)* detects breaking news but it can be applied to other domain as well. Online clustering is based on similarity functions upon the content through a modified version of Lead-follower algorithm proposed by Shah (2010) which allows for clustering in both content and time. It aggregates tweets in clusters according to the topic they referred to, and the geographical area mentioned in tweets. In particular, they represent news tweet *t* with i feature vector representation using TF-IDF and compute the distance between *t* and a candidate cluster *c* using a variant of the cosine similarity (COS) measure. They modified the latter cosine distance in order to involve temporal dimension by applying the Gaussian attenuator. The difference in days between the cluster's mean publication time of all the tweets T_c and the tweet's publication time T_t, are taken into account in the online clustering as defined below:

$$\dot{\delta}(t,c) = \delta(t,c) \cdot e^{\frac{-(T_t - T_c)^2}{2(\sigma)^2}}$$

To distinguish relevant tweets from spam a naive Bayes classifier was trained with a manually built lexicon with keywords extracted from news articles published around the same period as tweets. The system also provides a UI with the news ranked in an order of importance and a map showing the geographic region of interest.

Table 9: Topic Detection and Tracking Tools

Tool	**TwitterStand** Sankaranarayanan et.al (2009)	**TwitterMonitor** *Mathioudakis & Koudas (2010)*	**Cloud4Trends** *Vakali et.al (2012)*	**TwiCal** Ritter et.al (2012)	**TweCom** Cagliero and Fiori (2013)
Trend Detection Service	Breaking News	General Topics	Local General Topics	Events	Spatial& Temporal Propagation analysis of trends
Techniques	-Naive Bayes - Document Based	-Clustering based on co-occurrences	-Online clustering	-Sequence Label with Conditional Random Field	Association Rule Mining;
	-Online clustering		-Feature Based	-Bayesian Model	Hierarchical Clustering; Semantic Ontologies; Genio & RuleGen; Feature Based;
	- Feature Based	-Feature Based	-Document Based	-Document Based	
Real-Time		YES	YES		
U.D.T.		YES	YES	YES	YES
S.D.T.	YES			YES	
Additional Features	Interactive UI with the concepts aggregated and geographically presented	Interactive UI with description for each trend	Capture user's trend history & geolocation	Group events into concepts including time & location of each event	Crawler to retrieve linked tweets and most significant trends
Similar to	Newsstand	Blogscope	TwitterStand TwitterMonitor	-	CAS-Mine
OSN	Twitter	Twitter	Twitter	Twitter	Twitter
Contribution Field	Online Clustering Geospatial Analysis	Burst detection and clustering algorithms	Cloud Infrastructure	Open-domain event extraction	Data Mining & SNA

U.T.D.: Unsupervised Topic Detection,
S.D.T. Supervised Topic Detection

Zhou et.al (2015) propose an end-to-end framework for filtering and categorizing events into concepts that also provides the location and time for each event. They filter events with two approaches: i) a keyword based through a lexicon which built manually in the same way as *TwitterStand, and* ii) a binary classification problem with features of frequently occurred words and patterns in event-related tweets. For extraction and categorization of events, they propose a simple Bayesian modeling (LECM) approach which is able to directly extract event-related keywords from tweets without supervised learning. Events in the framework are represented as a 4-tuple <y, d, l, k>, where y stands for non-location named entities, d for a date, l for a location, and k for event-related

keywords. It is assumed that in the model, each tweet message m is assigned to one event instance e, while e is modeled as a joint distribution over y, d, l and k. Their work is similar to TwiCal (Ritter et.al (2012)) in the sense that they also focus on the extraction and categorization of structured representation of events from Twitter. However, TwiCal relies on a supervised sequence labeler based on Conditional Random Fields and trained on tweets annotated with event mentions for the identification of event-related phrases. Whereas here all the methods are unsupervised and additionally an enhanced version of filtering is implemented. Both tools use POS, NER and temporal resolution to process tweets. Future work could be the use of cloud in order to reduce the error propagation resulted from the separate computation of the steps.

*Wang et.a*l (2016) study the problem of detecting events instead of fixed, in adjustable time windows. For instance, their system gives data scientist the ability to know about how a hot event, happened in the last 120 minutes, how it was developed during the 60, 30 and 10 minutes. To detect events, they use unigrams as terms for each new tweet, claiming that unigrams out-performs n-grams in both effectiveness and efficiency. They detect events through anomaly detection, namely they process each new tweet and store their statistics (number of retweets, number of tweets per minute, number of users and number of different retweeted users) and identify abnormal terms at the end of each current time window. The clustering is based on co-occurrences and the selection is based on the top-k ranked clusters. They design a data structure to support adjustable time window based event detection. Their proposed technique outperformed TwiCal in accuracy.

Politwi (Rill et.al (2014)) is a tool available on twitter, website and smartphone apps for detecting the top political German discussions in tweets hourly and daily. The hashtags are the base for topic detection and the emoticons contained in hashtags are the base for sentiment analysis. The basic idea of their TDT approach is to compare the current number of tweets with hashtag to the number of tweets of the previous period taking into account the standard deviation using the Gaussian distribution. To this extent, a top topic is characterized by a significantly higher current appearance compared to a previous time period. A graph is built with each hashtag (node) to be surrounded by links of the connected words used in the current context together with a predicted polarity for each one. The relation graph contains the most frequently occurred words in specific time

points and can be used to extend the existing knowledge bases for answering questions like “Which polarity bears the upcoming topic '#Merkel' in this political context?”.

*Sociopedia (*Kaushik et.al (2016)*)* is a different system for analyzing social media topics. It constructs automatically a semantic ontology based on a given keyword. The nodes in ontology are entities extracted from the retrieved top tweets and the relationships are inferred through related-documents form Wikipedia and DBpedia. POS and NER are implemented to construct the ontology as well. Since researcher's objective is to monitor a marketing campaign for a new product launch in Twitter landscape, the system includes a query summarization analysis, a comparison detection and a sentiment analysis component as well. The sentiment analysis conducted through the lexicon AFINN and the other two components are built through frequency distribution of word patterns. To illustrate the latter, the presence of the word ‘versus’ may indicate a comparison and the presence of 5W1H(what-where-who-why-whether-how) is an indicator of a query.

Table 10: Topic Detection and Tracking Tools (2nd part)

Tool	**Politwi** *Rill et.al (2014)*	***Zhou et.al*** (2015)	**Sociopedia** Kaushik et.al (2016)	***Wang et.al*** (2016)	**Nguyen and Jung** (2016)
Trend Detection Service	Political Topics in German	Events	Specific Events	Events in Adjustable time windows	Events
Techniques	-Statistical Analysis -Feature Based	-SVM Classifier -Bayesian Model -Document Based	-Semantic Ontologies -Statistical Analysis	-Clustering based on co-occurrences -Feature Based	-Density-Based Spatial Clustering -Online behavioral analysis -Feature Based
Real-Time	YES				YES
U.D.T.		YES		YES	
S.D.T.					
Additional Features	Website, Smartphone app &Twitter Representation	Group events into concepts including time & location of each event	-Lexicon-based Sentiment analysis -Query detection & Summarization -Comparison Detection	Modify the segment Tree Data Structure	-Diffusion Speed -Participants Number
Similar To	Google Trends	TwiCal	-	TwiCal	-
OSN	Twitter	Twitter	Twitter	Twitter	Twitter
Contribution Field	Big Data Apps	Unsupervised categorization of tweets	Business Intelligence App	Not fixed time windows	Online behavioral analysis

6.2 SUMMARY

A TDT system aims to "search, organize and structure" textual materials, mostly posts, from social networks and to answer the question of "what-where-when and by whom are the topics being set?". Some tools answers "what-where" questions like Cloud4trends, other focus on "what-when" like TwitterStand and other tools like that of Nguyen and Jung (2016) answers "what-by whom and when" but no one is able yet to handle the above research question completely.

Besides that, there are many other differences between the tools; some aim to help data scientists (TweCom), other aim to inform the end user(TwitterStand); some return a set of documents (TwitterStand) as trends and other keywords (TwitterMonitor); some focus on detecting specific-concept (TwiiterStand) whereas other are open-domain tools (Cloud4Trends) and lastly some tools support visualization (Politwi); some perceive detection in real-time **(**Cloud4Trends**)** and other in batch **(**TweCom**)** and some tools provide extra analysis components (Sociopedia).

Last but not least, semantic annotations such as named entities, geographic locations, and temporal expressions can help the system to extract topics.

7 SENTIMENT ANALYSIS (SA)

This chapter describes up-to-date SA frameworks developed to derive sentiments from OSN. An overview is given about the correlation among the methods used, the frameworks' research purpose and the text granularity they focus. All metrics are based in the theoretical baseline defined in sections 2.5.1 & 2.5.3. The results among the several tools are illustrated in Tables 11 and 12.

7.1 SA FRAMEWORKS

Tables 6 and 7 present a categorization of SA tools according to many parameters such as the year they were created, the type of sentiment analysis service they offer, the SA approaches and techniques they employ and whether the tools support concept level analysis. Specifically:

1. "Year" refers to the year the tool was created.
2. "Sentiment Analysis Service" demonstrates the service provided by the tool.
3. "Techniques" shows the specific analysis techniques used to develop the tool.
4. "U.M.L." and "S.M.L." stands for unsupervised and supervised machine learning methods respectively. A tool assigned into these classes entails that a statistical approach is utilized. The specific machine learning algorithms that used are mentioned into the "Techniques" class.
5. "Lexicon-based" indicates that the tool employed the lexicon based approach. The cell refers to the knowledge base, dictionary or manually built corpus that was utilized. When a tool is both assigned in "U.M.L." and/or "S.M.L." and "Lexicon-based" means that a hybrid approach is employed.
6. "Document" and "Sentence" refer to whether the tool identifies a sentiment at the sentence or document level.
7. "Concept" refers to concept-level sentiment analysis; which focuses on a semantic analysis of text through the use of web ontologies or semantic networks, which allow the aggregation of conceptual and affective information associated with natural language opinions [59].
8. "OSN" denotes the OSN each tool built for and tested on.
9. "Other Tools" points out the synergy of other tools within the framework in order

to achieve its analysis service.

A development environment for sentiment classification of tweets is proposed by Sheela 2016, as illustrated in Figure 18. She took into account the fact that processing and analysis algorithms should be aligned with the strict constraints of storage and time since user generated content arrive at high frequency and volume. Specifically, sentiment analysis is done in MapReduce layer via a Naive Bayes classifier and the results are stored in MongoDB.

COMPONENTS	ROLES
Operating System	Use of Hadoop for distributed storage Supporting Java environment for processing some business logic
Crawler, HDFS Layer	Crawler: Gathering the source data from various SNSs HDFS: Distribution File system, Data storage
MapReduce Layer	Sentence Analysis, Text Mining, Sentiment Analysis
MongoDB	Storing analyzed results by MapReduce in MongoDB
Web Server	Supporting Web applications using analyzed results

Figure 18: Development Environment for sentiment classification of tweets

Inspired by the coarse grained Machine learning analysis that treat each tweet as one uniform statement, *Kontopoulos* et.al 2013 utilize ontology-based techniques. In particular, they broke down each tweet into a set of features relevant to a pre-defined domain to give a more detailed analysis of posted opinions. They created an ontology with concepts and relations through the manual Formal Concept Analysis (FCA) methodology combined with the semi-automatic ontology editor Onto-gen. They enrich the domain ontology with synonyms using WordNet and they extract tweets relevant to the ontology's concepts. Lastly, they extracted sentiment from isolated sentences through a web service called OpenDover. FCA is a mathematical data analysis theory utilized to derive a hierarchy of concepts where each concept represents the set of objects (iphone) sharing the same values for a certain set of attributes (camera). A formal context is defined as a triple of:

$$K = (G, M, I)$$

where G is a set of objects, M is a set of attributes, and $I \subseteq G \times M$ is a binary relation expresses which objects have which attributes.

Poria et.al (2014) also utilized ontology-related technologies for concept level analysis but instead of creating a knowledge representation through mathematical logic, they applied semantic relationships with the support of SenticNet. They focus on augmenting the sentic computing framework with dependency-based rules that leverage syntactic properties of text for sentence-level polarity detection. The sentic computing framework, introduced by Cambria et.al (2010), process natural language via common sense tools and affective-semantic ontologies, besides mathematical and social concepts. Poria et.al contributed in better understanding of the contextual role of each concept within a sentence by allowing sentiments to flow from concept to concept based on the dependency relation of the input sentence. Dependency relations are useful in finding relations (links) between subjective words and a topic (Mithun(2012)). In particular, natural language text is first deconstructed into concepts, through POS and syntax analysis, to be later fed to a vector space of common-sense knowledge. The latter structures words in terms of their affective valence. It is built to analyze the concepts by means of an emotion categorization model (Hourglass model), which is inspired by human emotions and brain activity theories. The model can potentially synthesize the full range of emotional experiences in terms of just four emotions: Pleasantness, Attention, Sensitivity, and Aptitude and predict polarity of opinions, according to the formula:

$$p = \sum_{i=1}^{N} \frac{Pleasantness(c_i) + |Attention(c_i)| - |Sensitivity(c_i)| + Aptidute\ (c_i)}{3N}$$

where c_i is an input concept, N the total number of concepts, and 3 is the normalization factor since the Hourglass dimensions are defined as a float $\in [-1, +1]$.

Analogical reasoning on the semantic and affective relatedness of natural language concepts succeeded via ELM and SVM which cluster the vector space model with respect to the Hourglass model. ELM are ANN with a single hidden layer whose first weight matrix need not be tuned so "only learns the last layer". It works for generalized single-hidden layer feedforward networks (SLFNs) whose typical structure shown in Figure 19. The ELM learning problem settings require a training set, X, of N labeled pairs, where (x_i, y_i), where $x_i \in \mathcal{R}^m$ is the *i*th input vector and $y_i \in \mathcal{R}$ is the associate expected 'target' value. The input layer has m neurons and connects to the 'hidden' layer (having O_h neurons) through a set of weights $\{\widehat{w}_j \in R^m; j = 1, \dots, O_h\}$. The *j*th hidden neuron embeds a bias term, $\widehat{b}_j$, and a nonlinear 'activation' function, φ(•); thus, the neuron's response to an input stimulus, x, is:

$$a_j(x) = \varphi(\widehat{w} \cdot x + \widehat{b}_j)$$

The overall output of the network is:

$$f(x) = \sum_{j=1}^{O_h} \overline{w}_j \, a_j \, (x)$$

To train an ELM involves the following steps:

1. Randomly set the input weights $\widehat{w}_j$ and bias $\widehat{b}_j$ for each hidden neuron;
2. Compute the activation matrix, H such that the entry $\{h_{ij} \in H; i=1,\ldots,o; j=1,\ldots,o_h\}$ is the activation value of the jth hidden neuron for the ith input pattern. The H matrix is:

$$\begin{bmatrix} \varphi(\widehat{w}_1 \cdot x_1 + \widehat{b}_1) & \cdots & \varphi(\widehat{w}_{O_h} \cdot x_1 + \widehat{b}_{O_h}) \\ \vdots & \ddots & \vdots \\ \varphi(\widehat{w}_1 \cdot x_O + \widehat{b}_1) & \cdots & \varphi(\widehat{w}_{O_h} \cdot x_O + \widehat{b}_{O_h}) \end{bmatrix}$$

3. Compute the output weights by solving a pseudo-inverse problem as:

$$\overline{w} = H^{+} y$$

In addition to that, the ELM was trained to act as a reserve to detect the polarity of the sentence when the concept wasn't found in the priori polarity lexicon of concepts SenticNet or didn't found any sentic patterns for it. In order to compute polarity, sentic patterns leverage on the SenticNet framework and on the syntactic dependency relations found in the input sentence. Although, the proposed approach is tested on offline datasets of movie reviews and electronics product reviews, we could say that the approach could be applied to posts from OSN with some crucial modifications.

It is worth noting that since the accuracy of the system crucially depends on the quality of the output of the dependency parser, which relies on grammatical correctness of the input sentence, a thorough cleaning and preprocessing part should be managed so as not to penalize results, since OSN's do not have predictable discourse structure.

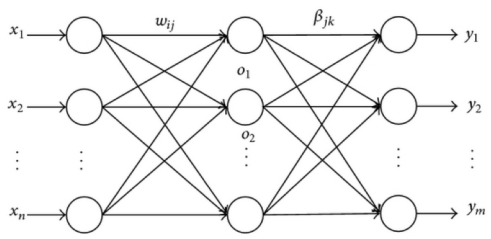


Figure 19: Single hidden layer feedforward networks

Due to the fact that the majority of such state-of- the-art frameworks rely on processing

a single modality, i.e., text, audio, or video, another work of Poria et.al (2016) propose a system for multimodal sentiment analysis from videos posted on YouTube, as illustrated in Figure 20. They extracted facial expressions features with ELM, vocal intensity features from audio track and they extracted concepts from texts following the sentic computing paradigm.

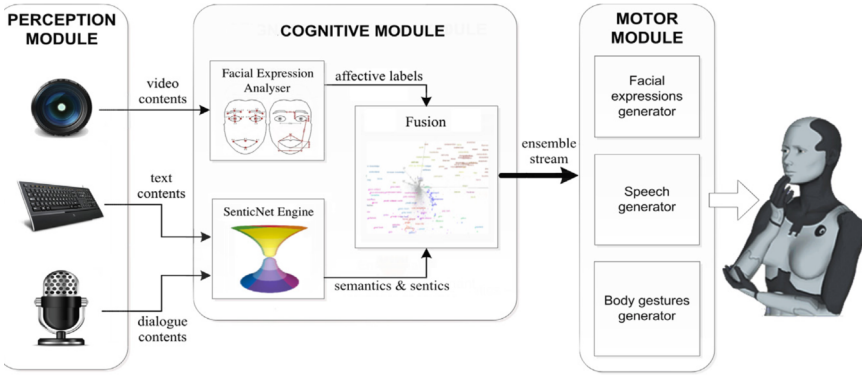


Figure 20: Multimodal SA on YouTube by Poria et.al(2016)

Also Yu et.al (2016) analyze sentiments based on multimodality content and specifically they employed deep learning models to extract both textual and visual features to analyze the sentiment expressed in Chinese microblogs. They adopt Deep Convolutional Neural Networks (DNNs), which are really popular in image recognition, with DropConnect to learn visual features. Also they trained another Convolutional Neural Network(CNN) using the short text of microblogging platform, and then, predicted sentiments by combining these two results. CNN is a feedforward network whose connectivity pattern between its neurons follows the organization of the animal visual cortex. Here, it is trained on a pre-trained word vector of Chinese characters resulted from word2vec tool. The word2vec tool is a set of neural networks that takes a text corpus as input and produces the word vectors as output. It first constructs a vocabulary from the training text data and then learns vector representation of words.

Mithun (2012) proposed a query-based opinion summarization framework called Blogsum that given a query and a set of blogs generates a summary of opinions. To extract and select the initial candidate sentences for the summary, BlogSum ranks each sentence using the features shown below:

$$\text{Sentence Score} = w1 \times \text{Question Similarity} + w2 \times \text{Topic Similarity} + w3 \times \text{Subjectivity Score}$$

where, question similarity and topic similarity are calculated using the traditional

technique of TF-IDF. The subjectivity score is calculated using a dictionary-based approach MPQA and is defined as:

$$Subjectivity\ score\ of\ a\ sentence = \frac{sum\ of\ the\ polarity\ score\ of\ all\ subjective\ words\ found\ in\ the\ sentence}{totalnumber\ of\ subjective\ words\ in\ the\ sentence}$$

They used heuristics rules to select the best possible sentences that would generate the summary. Similarly, with Poria et.al 2014, they take advantage of dependency rules in order to identify whether the sentence topic is associated with any of the subjective words of the sentence. Though, they use different methods and they included other discourse relations as well. For instance, their system identifies whether a comparison is present within a sentence via Naive Bayes and class sequential rule classifiers.

Prom-on et.al (2016) also included a summarization component in their opinion mining framework. The framework is composed by two other analysis components: sentiment analysis and the influencer analysis. The latter is done via measuring the degree centrality of the network. The out-degree is defined as:

$$C_{D_o}(i) = \sum_{j=1}^{n} a_{ij}$$

where i is the user and aij the relationship with another user, aij =1 when a user follows someone otherwise is equal to 0. Respectively at the in-degree defined below, aji=1 when the user is followed by someone otherwise is equal to 0.

$$C_{D_t}(i) = \sum_{j=1}^{n} a_{ji}$$

For the SA they defined five corpora with syntactic features, positive and negative words to employ a lexicon-based algorithm that iteratively matches sentiment keywords with the remaining corpuses. The summarization analysis is created through sentence similarity calculation, sentence clustering and last sentence selection. The sentence similarity score is defined with the usage of a predefined similarity word-pair corpora and the vector space model. Each sentence is represented with a set of words and the merge of two word sets of the two comparing sentences, supports two semantic vectors (V1, V2) to be created. Each element of vector represents the similarity score between a word pairs (r,s) in similarity corpora, as we illustrate in Figure 21.

Finally, the similarity score between two sentences is derived from a cosine similarity cos $\theta \in [0,1]$ between V1, V2 as defined

$$cos\ \theta = \frac{V_1 \cdot V_2}{||V_1|| \cdot ||V_2||}$$

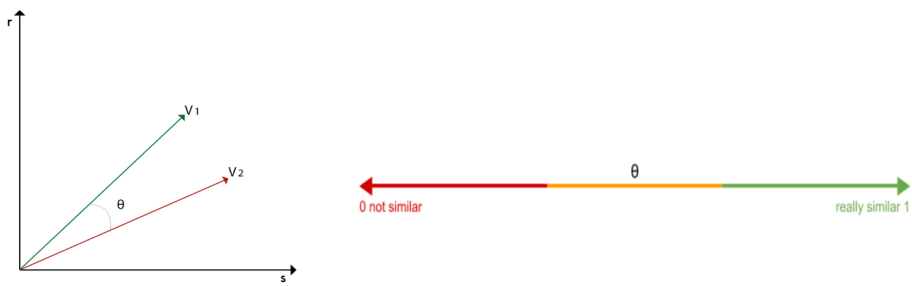


Figure 21: Similarity calculation between two sentences in vector space text representation

Then they minimize the dissimilarity between sentences in the clusters through a modified version of AI's Genetic Algorithm so as to eventually select the underlying text from each cluster. The modification is that they take into account the membership degree of a sentence in the cluster to boost up the results.

Similarly, Li and Li (2013) propose a keyword-based framework for numeric sentiment summarization, accompanied also with influencer analysis to evaluate the credibility of a user. The credibility score of user i in a social network SN in a time period TP and obtained by:

$$CS_i = \sqrt{f_i^{SN} \times r_i^{TP}}$$

where

f_i^{SN} is the credibility of an expresser and defined as:

$$f_i^{SN} = \frac{\sum_{j \neq i}^{N} SN_{j,i}}{\sum_{j \neq i}^{N} SN_{i,j}}, 1$$

N is the users in the social network SN. If user i follows user j then $SN_{i,j}=1$, otherwise $SN_{i,j}=0$.

r_i^{TP} is the credibility of the content and defined as:

$$r_i^{TP} = \frac{number\ of\ posts\ reposted\ of\ user\ i\ in\ time\ period\ TP}{number\ of\ posts\ of\ user\ i\ in\ time\ period\ TP}$$

Features/topics were extracted via TF-IDF keyword frequency multiplied with meronym pattern $MPP_{q,t}$ which is defined as:

$$MPP_{q,t} = \frac{number\ of\ occureneces\ of\ t\ in\ O_q\ with\ pattern\ P}{TF_{q,t}}$$

where P is a set of predefined meronym patterns, q is the given keyword, t is a distinct term in a set of phrases/nouns and O_q is the set of tweets containing term q.

To score subjectivity of opinions, a subjective word set, Φ, was built with WordNet and the opinion subjectivity for a post o related to a topic t, $OS_{O,t}$, formulated as:

$$OS_{o,t} = (\sum_{s \in S_t^o} \frac{|U_s \cap \Phi|}{U_s})/|S_t^0|$$

where S_t^O is the set of sentences in opinion o which mentions topic t, and U_S is the set of unigrams pertained in sentence.

To classify the polarity of opinions, an SVM classifier used, trained on emoticons. The semantic score of opinion o on a topic t is defined as $SS_{o,t}$ whose values range between 1 for positive and -1 for negative sentiments.

$$SS_{o,t} = Polarity_o \text{ x } OS_{o,t}, \text{ where } SS_{o,t} \in [-1,1]$$

The final score for a topic t with respect to a query q is formulated as

$$Score_{q,t} = \frac{\sum_{o \in O_{q,t}} (SS_{o,t} \times CS_i)}{\sum_{o \in O_{q,t}} (|SS_{o,t}| \times CS_i)}$$

and $O_{q,t}$ is the set of opinions mentioning topic t for a given query q and user i is the expresser of an opinion o.

Cuesta et.al (2014) propose a customizable and extensible framework to analyze sentiments of Spanish tweets and generate reports as output. A language-agnostic sentiment analysis module provides a set of tools to perform sentiment analysis. In particular, polarity classification, performed manually through NLTK Python interface, in order to create corpora for training data. The system classifies texts, given the Naive Bayes classifier and a set of ngrams, and the most probable category is returned (either "positive" or "negative") along with its probability.

Table 11: Sentiment Analysis Frameworks

Tool	Kontopoulos et.al(2013)	Mithun(2010) Mithun(2012)	Li and Li (2013)	Prom-on et.al (2016)	Cuesta et.al (2014)
Sentiment Analysis Service	Feature-based SA on specific topic	Opinion Summarization given a Query&Blogs	Keyword-based Numeric Opinion Summarization with Influencer Analysis	Keyword-based opinion mining in Thai Language accompanied with a summary and influence analysis	Sentiment Classification and quantitative analysis of Spanish Tweets
Techniques	Ontology-Related FCA	Discourse Relations Heuristic Rules Naive Bayes Sequential Rules TF-IDF	TF-IDF & Meronym patterns SVM on emoticons	Genetic Algorithm Cosine Similarity	Naive Bayes
U.M.L.	YES		YES		
S.M.L.			YES		YES
Lexicon-based	WordNet	MPQA	WordNet	Manual Corpus	Manual Corpus
Document			YES		YES
Sentence	YES	YES	YES		
Concept	YES	YES			
OSN	Twitter	Blogging Services	Twitter	Twitter Facebook Foursquare	Twitter
Other Tools	Ontogen OpenDover	Spade Parser Stanford POS	Hits&Pagerank	MapReduce MongoDB	MapReduce Python NLTK NodeJs

Knime is an open platform for analytics fitting in big data era. It is used together with Hadoop and HBase to analyze sentiments regarding specific brands presented on online reviews and Twitter by Minanovic et.al (2014). Researchers analyzed tweets at word level via a lexicon based approach MPQA and they counted the significance of each word via TF-IDF. Researchers showed that sentiment analysis of online reviews is less complicated process but more time and resource intensive; whereas a vice versa situation is observed in tweets sentiment analysis.

A different method from the above, presented by Hu et.al (2013) which took advantage the networked nature of posted messages through user connections, including both user-message and user-user following relations. They employ the unigram model to construct the feature space and use term presence as the feature weigh. They transform user-centric social relations into sentiment relations between tweets based on social theories of Sentiment Consistency and Emotional Contagion. The basic idea is to build a latent connection, mathematically formulated, to make two messages as close as possible

whether they are posted by the same user (Sentiment Consistency) or they are follower/friend with each other (Emotional Contagion). To retain original information in the texts and discard the noise, instead of term filtering through dictionaries, they model the relations using graph Laplacian, which is employed as a regularization to a sparse formulation. After modelling the above sentiments relations, the sentiment classification of microblogging data formulated as the following optimization problem:

$$\min_{w} \frac{1}{2} ||X^T W - Y||_F^2 + \frac{a}{2} ||W^T X \mathcal{L}^{\frac{1}{2}}||_F^2 + \beta ||W||_1$$

where T = [X, Y]: X is the content matrix, Y is the sentiment label matrix and T the given corpus with messages. W is the desired classifier to automatically assign sentiment labels for unseen messages (i.e., test data), α and β are positive regularization parameters and F is the user-user matrix and last but not least $\mathcal{L}$ is the Laplacian matrix for the different message-message relations.

Table 12: Sentiment Analysis Frameworks (2nd part)

Tool	Sheela et.al 2016	Poria et.al 2016	Hu et.al 2013	Yu et.al 206	Poria et.al 2014
Sentiment Analysis Service	Development Environment for sentiment classification	Generic Multi-modal SA system	Model social relationships between tweets	Polarity detection engine combining both images and Chinese text	Sentence Polarity detection engine using dependency rules & Sentic Computing
Techniques	Naive Bayes	ELM	Regression Analysis Least Square	CNN DNN	ELM SVM Ontology related and semantic Technologies
U.M.L.		YES		YES	YES
S.M.L.	YES		YES		
Lexicon-based	SenticNet WordNet	SenticNet EmoSenticNet			SenticNet AffectNet
Document			YES		
Sentence	YES	YES		YES	YES
Concept		YES			YES
OSN	Twitter	Youtube	Publicly Available Twitter Datasets	SinaWeibo	Movie Review & Blitzer datasets
Other Tools	Hadoop MapReduce MongoDB	Matlab Luxand FSDK 1.7 GAVAM OpenEAR Sentic Computing	Stanford Twitter SentimenT&ObamaMcCain Debate Social Theories Laplacian Matrix	Word2vec Python jieba DropConnect Sentic Computing Logistic Regression	Stanford Chunker Hourglass Model Analogical Reasoning AI

7.2 SUMMARY

Analysis of sentiments via mere word-level analysis namely BoW and subjective words counting extract results that are simply not true. Since they assume the independence of words and ignores the importance of semantic and subjective information in the text. A new frontier in sentiment analysis is developing context aware systems that detect the changing of meaning in changing contexts. Besides, incorporating network features, another approach to this issue is the recent movement towards concept-level sentiment analysis. The latter focuses on a semantic analysis of text through the use of web ontologies or semantic networks, which allow the aggregation of conceptual and affective information associated with natural language opinions. Except for the bag-of-concepts methods, recent prominent computational intelligent methods like deep learning are also

popular for their ability to analyze the sentiment of short texts by learning sentiment representations from a large corpus of labeled and unlabeled text. Hence, these recent techniques should be utilized in order to new results to be drawn and create a more profound theoretical framework.

8 COLLABORATIVE RECOMMENDATION(CR)

This chapter describes CR frameworks designed to capture different user behavioral models and types of available information. A correlation between the methods used in CR frameworks and the recommendation task they provide is also given. All metrics are based in the theoretical baseline defined in section 2.5.4. The results among the several tools are illustrated in Table 13.

8.1 CR FRAMEWORKS

Collaborative recommendation is effective at representing a user's overall interests and predicting accurate recommendations to users according to their preferences.

Hsu et.al (2012) proposed a personalized auxiliary learning material recommendation system using Facebook searching queries. The system recommends to learners appropriate learning items that are both best match to the query and are the most people's likeable. It also takes into account attributes like the degree of difficulty of the auxiliary materials, individual learning styles and the specific course topics. Authors implement population-based optimization algorithm called the Artificial Bee Colony (ABC) to optimize the results of recommendation.

Khalid et.al (2014) presented a cloud-based framework for context-sensitive venue recommendations in social networks for a single user or a group of friends with similar interests. The system uses the opinion of experienced users to recommend items. It combines CF with the group satisfaction principle of social computing. After ranking users and venues in a geographic location via HITS mechanism (item-based technique), they create a similarity graph among a set of experienced users (called hubs) who share the similar preferences for various venues (user-based technique). Then, they apply a variant of ant-colony algorithm to generate an optimal venues selection that best match user's preferences. To deal with the scalability problem, they use cloud infrastructure.

Chen et.al (2016) propose a recommendation scheme to make followee recommendation in Twitter leveraging implicit information. They compute similarity between users via a modified latent factor model with top-k ranking optimization criterion in reasons that

conventional similarity functions face high scalability issues. After computing all the preference pairs of followee candidates for a given target user *u* and the rank order of a followee candidate *i* in *u*'s preference list, latent features of tweet content factors and social relationship factors (frequency of retweeting or placing comments to another user) are incorporated to recommendation system. To combine tweet users post, the factorization model is defined as:

$$y_{u,i} = bias + p_u^T (\frac{1}{\beta} \sum_{w \in T_i} t_w)$$

where bias is used to indicate any form of possible bias to simplify the equation; T_i is the term set of tweets posted by user i; β is the normalization term for features and t_w corresponds to a certain term vector mentioned by user i.
To incorporate social relations a model is defined as:

$$y_{u,i} = bias + p_u^T S_i$$

where s_i is the latent factor of the followee candidate i.

He and Chu (2010) leverage OSN's explicit social interactions to design a probabilistic model which makes recommendations based on user's own preferences, the general acceptance of the target item, and the opinions from social friends. Specifically, they use influence analysis and select an appropriate set of friends according to the type of target items, known as semantic filtering. They incorporate the influences from both distant friends and immediate friends inference that calculated through a Naïve Bayes classifier; Immediate friends are considered those who are just one hop away from each other in a social network graph and distant friends are those who are multiple hops away. The core of the recommender system is to predict the probability distribution of the target user *U*'s rating on the target item *I* given the attribute values of item I, the attribute values of user *U*, and the ratings on item *I* rated by *U*'s immediate friends as defined:

$$\Pr(R_{UI} = k \mid A' = a'_I, A = a_U, \{R_{VI} = r_{VI} : \forall V \in U(I) \cap N(U)\})$$

Sun et.al 2015 propose a matrix factorization framework that combines both friendships and ratings records as social regularization terms. Researchers combine the friendships (immediate) among users selecting the the most 'suitable set of friends with a biclustering algorithm'. They calculate similarity between users taking into account tags they share as defined below:

$$w_{ut} = \sum_{i} \frac{1}{|M_{ui}|} \; if \; t \in M_{ui} \;\; i \in I$$

where w_{ut} denotes the weight of tag t labeled by user u, M_{ui} is the tag list that user u gave to item i, and $|M_{ui}|$ is the number of tags. Then they compute correlation between users and items via mapping them in the tag space model as defined:

$$w_{jt} = \sum_{u} \frac{1}{|M_{uj}|} \; if \; t \in M_{uj} \;\; u \in U$$

where w_{jt} denotes the weight of tag t of item j, M_{uj} is the tags list and $|M_{uj}|$ is the number of tags which user u gave to item j.

The similarity calculation for both of the two algorithms is cosine similarity. The general idea of matrix factorization (MF) is to model the user-item interactions with factors representing latent characteristics of the users and items in the system. The model is trained using data from Del.icio.us, and later used to predict ratings of users for new items.

A different approach presented by Sieg et.al (2010) who incorporate semantic knowledge from ontologies to enhance a context-sensitive collaborative recommendation. The centric idea is to create ontological user profiles that are learned and incrementally updated with the support of an underling ontology of concepts in a particular domain of interest. After calculating the users' level of interest for each concept, compare the ontological user profiles for each user to form semantic neighborhoods in order to compute the similarity among user profiles. The prediction of a user's rating for an item is calculated with a variation Resnick's standard prediction formula. The ontology relies on existing hierarchical taxonomies of Amazon.com's Book Taxonomy.

Although Frasincar et.al (2012) proposal neither is evaluated on social networks nor uses collaborative recommendation, it is interesting since it aims to propose previously unseen items, news items, by semantically expanding user profiles using ontology-related methods. They compute similarity relations both between concepts per se and between to concepts related to the user profile. They use a knowledge base whose quality directly influence the accuracy of recommendation results. Importantly, there are many future directions that should be checked and/or be solved such as the matter of the manual

maintenance of the knowledge base.

Table 13: Tools for Collaborative Recommendation

Article	Content -Based	Model-based	Memory-based	Techniques			Task	OSN
Chen et.al 2016	YES	YES		Latent factor model TF-IDF			Followee recommendation	SinaWeibo
Sun et.al 2015				Matrix factorization based on social regularization			Rating Prediction	De.li.cio.us
He and Chu 2010		YES		Influence Analysis	Naïve Bayes		Personalized Recommendation	Yelp
Hsu et.al 2012		YES		Artificial Bee Colony			Personalized Learning Material Recommendation	Facebook
Khalid et.al 2014		YES	YES	Cloud	HITS	Ant-Colony	Group and Personalized Venue Recommendation	Mobile Social Networks
Sieg et.al 2010	YES		YES	Ontology Related and Semantic Technologies			Personalized Book Recommendation	Book-crossing Community
Frasincar et.al 2012			YES	Ontology Related and Semantic Technologies	Knowledge-base		Personalized Recommendation of News	-

8.2 SUMMARY

Current work on CR systems in OSN's has demonstrated the effectiveness of incorporating social network information to improve recommendation accuracy. Generally, semantic computing like user profile construction or similarities of user interactivity with content and SNA like community extraction or identification of trustiness are blended inside nearest-neighbor approaches and model-based approaches. It is not hard to observe that model-based and specifically factorization approach is dominant in the most recent papers. The mere and classical form of nearest-neighbor approaches demonstrate ineffective performance. Lastly, computing Intelligence techniques have recently exhibited significant potentials to make recommender systems more robust, effective, and context-aware. Interestingly, Abbas et.al (2015) observed that

each of the CI technique is capable of dealing with one or more challenges and recommender systems may need to utilize the CI techniques in conjunction with each other to entirely deal with the challenges stated.

9 Open Issues and Potentials

This chapter demonstrates the challenges of techniques in social networking data analysis and proposes possible future directions.

9.1 Limitations of Machine Learning techniques

Machine learning is a subset of Computing Intelligence that mimic learning process that doesn't rely on linear logic of "if-then". Most event detection and sentiment analysis algorithms tackle the problem, at least in a first stage, as a clustering task either with supervised classifiers trained on textual features (e.g. n-grams) and structural features (e.g. number of followers) or unsupervised based on a scoring function to classify clusters. Semi-supervised learning exploits a small amount of labeled data together with the large amount of unlabeled data to build classifiers, however this approach is sensitive to classification efficiency and threshold settings, according to Atefeh and Khreich (2015).

In SA area has been showed that supervised approaches tend to overcome unsupervised ones (Musto et.al (2014)) because of classifiers automatically "learning" the task based on historical cases. Despite the low computational cost of the Naïve Bayes, it has not been competitive in terms of sentiment classification accuracy when compared to SVM (Li and Li (2013); Musto et.al (2014)). A comparison between SVM and Artificial neural networks (ANNs) is presented by Moraes et.al (2013) regarding document-level sentiment analysis and they indicate that ANN actually outperformed SVM on movie reviews.

However interesting machine learning techniques are, they have a lot of limitations. Classifiers do not work well at sentence level analysis since they require large text input. Moreover, they presuppose a training data set which isn't always available; especially in the case of OSN's there are many reasons such as privacy that restrict datasets to be public. Additionally, supervised techniques typically assuming OSN as a static environment. Unsupervised techniques, need scalability, optimization on setting the thresholds of incremental clustering algorithms that should be based on more adaptive rather than simply relying on static features. Effectiveness of both machine learning techniques rely on feature engineering, a labor-intensive and domain-dependent task.

Deep Learning algorithms are one promising avenue of research into the automated extraction of complex data features at high levels of abstraction (Najafabadi (2015)). One reason that justifies deep learning as a powerful way to analyze big data is that it allows computers to learn, without being taught, namely by avoiding human intervention.

9.2 Limitations of Lexicon-based techniques

While lexicon-based approach has become dominant within the field of text mining, it does have many limitations. Lexicon-based techniques either rely on an online dictionary, a knowledge base or on a manually labelled corpus. The latter is infeasible considering large-scale data and it is also subjective to biases. Online dictionaries and knowledge bases on the other hand, depend heavily on comprehensive knowledge representation and lack set of words and concepts. Being unable to find opinion words with domain and context specific orientations, knowledge bases are usually blended with the corpus-based approach which rely on syntactic or co-occurrence patterns of discourse structure. Though, the informal language of web-posts restricts this approach.

Sentic computing (Cambria and Hussain (2015); Cambria et.al (2010)) is a new paradigm to concept-level sentiment analysis that combines deep learning techniques with lexicon based ones to infer polarity from the text. Generally, the framework incorporates the following according to Poria et.al (2014) and Cambria and Hussain (2015):

- Artificial Intelligence and Semantic Web techniques, for knowledge representation and inference;
- Mathematics, for carrying out tasks such as graph mining and multi-dimensionality reduction;
- Linguistics, for discourse analysis and pragmatics;
- Psychology, for cognitive and affective modeling;
- Sociology, for understanding social network dynamics and social influence; and
- Ethics, for understanding related issues about the nature of mind and the creation of emotional machines.

This is a result from the move from traditional word-based approaches, towards semantically rich concept-centric approaches, combining both computer and social sciences together in order to endow machines the ability to learn things we know about

the world so as to better process natural language text (Davis and Marcus (2015)). Still, the framework leverages SenticNet knowledge base and since lexical resources are mostly in English, adopting them to other languages is considered a challenge.

9.3 Computational Intelligence in Social Networking Data Analysis

Importantly, data cleansing and analysis both borrow concepts and tools mostly from graph theory, text mining and conventional machine learning techniques. However, less attention has been attracted the Computational Intelligence paradigm.

CI encompasses algorithms like artificial neural networks (ANN), fuzzy systems (FS), evolutionary algorithms (EA), swarm intelligence (SI) and artificial immune systems (AIS) and all mimic procedures observed in nature. Owing to their promising property of adapting in a changing environment (Siddique and Adeli (2013)) of CI, they could be used in sentiment analysis systems that with a few changes at their core program they could work well in any language, not just in English. Other significant attributes CI poses are generalization, discovering, reasoning and association (Rambharose and Nikov (2010)), all beneficial to social networking data analysis.

EA's Genetic Algorithm is inspired by the Darwinian struggle for existence, where only the fittest individuals can survive in nature. It has found application in generating summaries from posts where only the fittest sentences should be selected. Genetic Algorithm and the Swarm Intelligence algorithms like Ant Colony and Artificial Bee Colony(ABC) are all population-based optimization algorithms. Ant Colony Algorithm imitates the ants' network of paths that connects their nests with the sources of food and ABC is inspired by the behavior of honey bees when seeking a quality food source. Both of the two have been used in personalized recommender systems using social media (Hsu et.al (2012)). ABC has recently shown promising results in clustering natural language morphemes (Sulaiman et.al (2015)).

As stated in 9.1, most previous research takes text analysis problems as a ranking task and employ learning-to-rank algorithms based on constructing novel features (e.g., lexical features, syntactic features, and semantic features), a time-consuming and labor-consuming problem which needs priori knowledge and usually a big dataset. In contrast

to more conventional machine learning and feature engineering algorithms, Deep Learning ANN has an advantage of potentially providing a solution to address the data analysis and learning problems as mentioned earlier since they go beyond mimic the learning process by imitating neural activation of human mind. The program is made of tangled layers of interconnected nodes and learns by rearranging connections between nodes after each new experience. One of the promises of deep learning is replacing handcrafted features with efficient algorithms for unsupervised or semi supervised learning and hierarchical feature extraction (Fu et.al (2016)). ANN are usually used to address classification and regression problems and actually apart from sentiment classification is used for social network classification as well (Perozzi et.al (2014)).

Understanding that the complex nature of human language isn't a machine understandable one, researchers should attempt to apply these techniques in OSN research. We summarize the meager social media analysis which is done through these techniques in Table 14. Though, to the best of our knowledge, in TDT weren't any application of CI.

Table 14: Computational Intelligence in Social Networking Data Analysis

Article	Technique	CI Category	Inspiration	Analysis Task	Objective
Yu et.al 2016	Convolutional Neural Networks	ANN-Deep Learning	Animal Visual Cortex	Learn Textual features	Classification
Yu et.al 2016	Deep Convolutional Neural Networks	ANN-Deep Learning	Animal Visual Cortex	Learn Visual features	Classification
Poria et.al 2016	Extreme Learning Machine	ANN-Single Hidden Layer Feedforward NN	Human Mind	Learn Facial Features	Classification
Poria et.al 2014	Extreme Learning Machine	ANN-Single Hidden Layer Feedforward NN	Human Mind	Classify polarity	Classification
Poria et.al 2014	Extreme Learning Machine	ANN-Single Hidden Layer Feedforward NN	Human Mind	Relate semantic and affective features of concepts	Regression
Prom-on et.al 2016	Genetic Algorithm	EA-Optimization	Evolution Process	Minimize dissimilarity between sentences	Clustering
Khalid et.al 2014	Ant Colony Algorithm	SI-Optimization	Ant's Food and Nests Network	Minimize dissimilarity between venues that best match to user preferences	Clustering
Hsu et.al 2012	Artificial Bee Colony Algorithm	SI-Optimization	Honey Bees searching quality food	Minimize dissimilarity between learning materials that best match to user query	Clustering

10 EXPERIMENTAL ENVIRONMENT

In this section, we detail the experiments conducted to verify the current level of sentiment prediction effectiveness using two machine learning API tools. To materialize the experiments, a cross-platform tool is built which is useful for sentiment classification of large sale data, through machine learning API services. The whole code required to build the tool can be found in: https://github.com/annishared/senti.

10.1 CLOUD-HOSTED MACHINE LEARNING SERVICES

There is a great variety of sentiment analysis API that given a text they provide its polarity through HTTP requests. They mainly achieve that, by sending the data to their pretrained machine learning models that are hosted on cloud. Essentially, the modelling part of ML has done in batch and the scoring of the sentences is occurring in real-time. Since an API is a pre-packaged solution for analyzing text, one doesn't need to build or train their own machine learning models. Importantly though, for more advanced researchers, Google has their own machine learning library TensorFlow5 which recently went open source and allows flexible models deployment through data flow graphs.
The two API's were selected in terms of (i) being recently deployed, (ii) allow to publish or perform any performance tests, (iii) provide no-charge API calls for academic research (iv) allows a decent number of API calls.

10.1.1 SENTIMENT CLASSIFICATION WITH GOOGLE NL & INDICO API

Google Cloud Natural Language API support sentiment analysis, syntax analysis and entity extraction in English language. On the other side, Indico API provides text analysis such as sentiment analysis, emotion categorization and entity extraction in a range of languages. Both take a sample of text from document, webpage or social media and return positive or negative sentiment which is represented by a numerical value that also involves the valence of sentiments.

10.2 DATA COLLECTION AND DESCRIPTION

To assess the performance of sentiment analysis methods over OSN's a small set of evaluation datasets has been released in the last few years. We have focused our selection on those datasets that are: (i) already classified tweets in terms of sentiments (ii) publicly available to the research community, (iii) manually annotated, providing a reliable set of

judgments over the texts, (iv) used to evaluate other sentiment analysis models and (vi) include sentiment labels: Negative and Positive which are not determined in terms of emoticons.

10.2.1 Dataset 1: STS-Gold Dataset

This dataset was built for the paper "Evaluation Datasets for Twitter Sentiment Analysis A survey and a new dataset, the STS-Gold" (Saif et.al(2013)). It contains 2034 tweets that were manually labelled by three graduate students. Although the dataset distinguishes between the sentiment of a tweet and the sentiment of entities mentioned within it, we utilize just the former one.

10.2.2 Dataset 2: Health Care Reform Dataset

The Health Care Reform (HCR) dataset was built by crawling tweets containing the hashtag "#hcr" (health care reform) for the paper "Twitter Polarity Classification with Label Propagation over Lexical Links and the Follower Graph" (Speriosu et.al(2011)). We select the subset of this corpus that was manually annotated by the authors with 3 labels positive, negative and neutral and contains in total 621 tweets.

10.2.3 Dataset 3: IMDB Reviews Dataset

This dataset contains 748 sentences from IMDB reviews, manually labeled by authors. It was created for the Paper 'From Group to Individual Labels using Deep Features', (Kotzias et.al(2015)) and include other reviews datasets as well.

10.3 Application Design and Implementation

Document NoSQL databases are suitable when one uses the same aggregate (document) to move back and front (Sadalage and Fowler (2012)). Since a document data model was needed, MongoDB, which is known for its scalability and flexibility, was chosen as a data store. After converting all datasets in JSON format, they were normalized in order to all have identical fields according to the model structure. A data classifier for each dataset was built in order to transform the numerical values of polarity to their corresponding string values.

Then, we built a tool in Node.js that executes the following processes, as are also illustrated in Figure 22:

1. Retrieves the documents from MongoDB.
2. For each record on the database, it sends a POST request of submitting the documents' content on the respective API.
3. Providing that, the tool successfully gets the sentiment of each document's content, it then updates the database with the specific sentiment classified with its sentiment class.

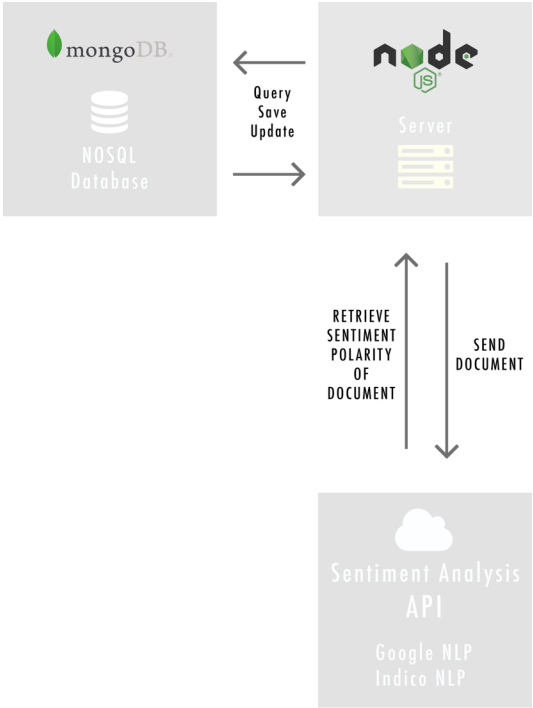


Figure 22: Application Architecture

The Node.js Server queries the database asynchronously (non-blocking) enabling it to communicate fast with an API which is highly required in real-time applications. This means that, when the response of the API is completed, it will write the result on the database without waiting for the response to happen. Node.js is really good at handling a lot of requests that are coming at once, like OSN.

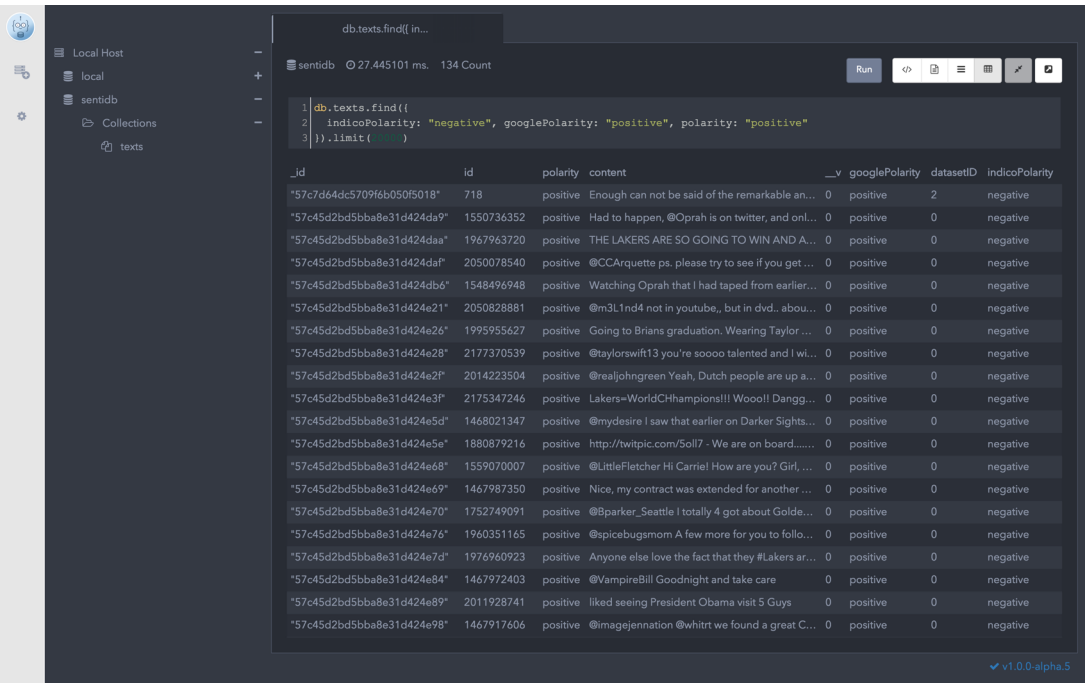
Besides the above, we use many state of the art technologies like mongoose and mongotron. The former is an Object Data Modelling (ODM) framework for a straight-

forward, schema-based modelling of the application data. Figure 23 demonstrates the data model defined for the datasets with the support of mongoose.

```
const mongoose = require("mongoose");
const Schema = mongoose.Schema;

/**
 * The Text schema we will use in our Text model
 */
const textSchema = new Schema({
  id: Number,
  polarity: String,
  content: String,
  datasetId: Number
});

/**
 * Create the Text Model based on the textSchema
 */
const Text = mongoose.model("Text", textSchema);


module.exports = Text;
```

Figure 23: Data Model for Datasets

The latter is a cross platform data management tool that handles the administration of MongoDB via both a GUI and code commands, as illustrated in Figure 24.



_id	id	polarity	content	__v	googlePolarity	datasetID	indicoPolarity
"57c7d64dc5709f6b050f5018"	718	positive	Enough can not be said of the remarkable an...	0	positive	2	negative
"57c45d2bd5bba8e31d424da9"	1550736352	positive	Had to happen, @Oprah is on twitter, and onl...	0	positive	0	negative
"57c45d2bd5bba8e31d424daa"	1967963720	positive	THE LAKERS ARE SO GOING TO WIN AND A...	0	positive	0	negative
"57c45d2bd5bba8e31d424daf"	2050078540	positive	@CCArquette ps. please try to see if you get ...	0	positive	0	negative
"57c45d2bd5bba8e31d424db6"	1548496948	positive	Watching Oprah that I had taped from earlier...	0	positive	0	negative
"57c45d2bd5bba8e31d424e21"	2050828881	positive	@m3L1nd4 not in youtube,, but in dvd.. abou...	0	positive	0	negative
"57c45d2bd5bba8e31d424e26"	1995955627	positive	Going to Brians graduation. Wearing Taylor ...	0	positive	0	negative
"57c45d2bd5bba8e31d424e28"	2177370539	positive	@taylorswift13 you're soooo talented and I wi...	0	positive	0	negative
"57c45d2bd5bba8e31d424e2f"	2014223504	positive	@realjohngreen Yeah, Dutch people are up a...	0	positive	0	negative
"57c45d2bd5bba8e31d424e3f"	2175347246	positive	Lakers=WorldCHhampions!!! Wooo!! Dangg...	0	positive	0	negative
"57c45d2bd5bba8e31d424e5d"	1468021347	positive	@mydesire I saw that earlier on Darker Sights...	0	positive	0	negative
"57c45d2bd5bba8e31d424e5e"	1880879216	positive	http://twitpic.com/5oll7 - We are on board.....	0	positive	0	negative
"57c45d2bd5bba8e31d424e68"	1559070007	positive	@LittleFletcher Hi Carrie! How are you? Girl, ...	0	positive	0	negative
"57c45d2bd5bba8e31d424e69"	1467987350	positive	Nice, my contract was extended for another ...	0	positive	0	negative
"57c45d2bd5bba8e31d424e70"	1752749091	positive	@Bparker_Seattle I totally 4 got about Golde...	0	positive	0	negative
"57c45d2bd5bba8e31d424e76"	1960351165	positive	@spicebugsmom A few more for you to follo...	0	positive	0	negative
"57c45d2bd5bba8e31d424e7d"	1976960923	positive	Anyone else love the fact that they #Lakers ar...	0	positive	0	negative
"57c45d2bd5bba8e31d424e84"	1467972403	positive	@VampireBill Goodnight and take care	0	positive	0	negative
"57c45d2bd5bba8e31d424e89"	2011928741	positive	liked seeing President Obama visit 5 Guys	0	positive	0	negative
"57c45d2bd5bba8e31d424e98"	1467917606	positive	@imagejennation @whitrt we found a great C...	0	positive	0	negative

Figure 24: Execute data queries on Mongotron

The tool is cross-platform and generic; but assumes that each dataset and API should comply with a specific form. It can handle large scale data efficiently; the only problem for a researcher would be the API's limitation requests. Extended work could include entity extraction or any other type of text analysis, usage of bigger datasets and detection of patterns that cause errors in the algorithm execution.

10.4 Experimental Results

For each evaluation dataset, both the level of precise sentiment prediction of each API was calculated and the number of their common inaccurate predictions.
Figure 25 depicts that Google API predicted a reliable number of tweets' sentiments. Almost the half of inaccurate results are commonly calculated by the two API's.

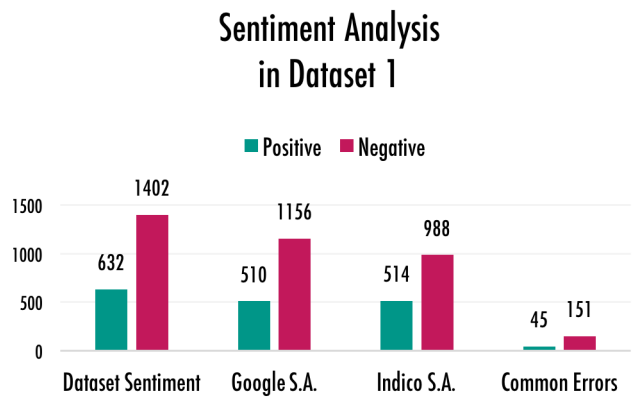


Figure 25: Sentiment Prediction Accuracy of Indico and Google in STS-Gold tweets

Figure 26 illustrates that each tool precisely recognized the sentiments of almost the half of total tweets. Google performance is more accurate on positive sentiments whereas Indico's is on negative ones. Again almost the half of inaccurate results are commonly calculated.

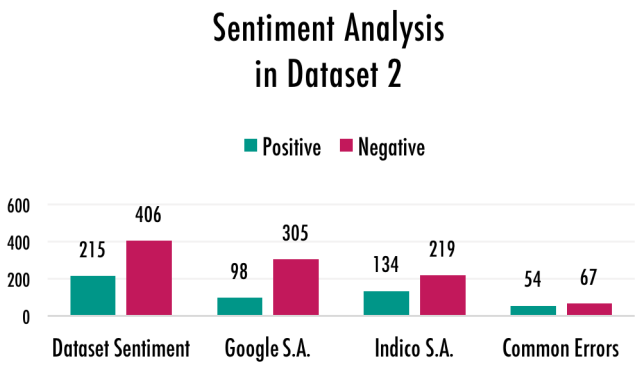


Figure 26: Sentiment Prediction Accuracy of Indico and Google in #hcr tweets

Figure 27 illustrates that both tools perform the same well and actually their most precise prediction among all datasets is observed here. This is quite reasonable since IMDB reviews contain human language that is more close to the norm of language structure and grammar.

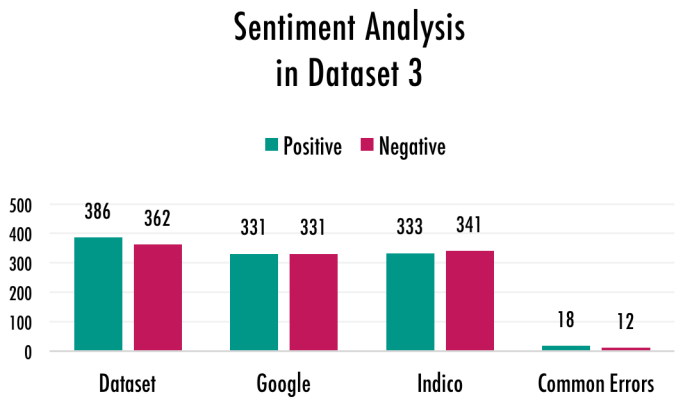


Figure 27: Sentiment Prediction Accuracy of Indico and Google in IMDB reviews

Figure 28 illustrates that accurate values in formal context reach a percentage of 89,3%, which is near to an excellent prediction; while a percentage of 70,5% is validly calculated in informal context. It is observed that informal OSN's posts, in our case tweets, are more difficult to be analyzed; since they are less structured than are the traditional blog articles, reviews or documents. Though, a great progress is noticed because the percentage of 30% inaccurate results is not considered a big one. However, what is not known is if the accurate predicted sentiments may be a result of randomness and not of correct calculations running behind. One way to discover that, is more datasets to be used, which

presupposes a greater number of API transactions to be allowed and more OSN polarity-tagged datasets to be available.

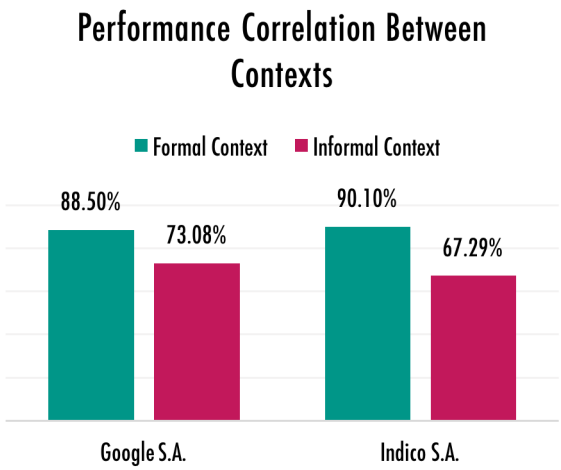


Figure 28: Sentiment prediction accuracy between formal and informal contexts

10.5 Summary

To summarize, each tool has its strengths and limitations but both performed really well which is an indicator that proves the progress of text analysis within the context of OSN. It is observed that, negative tagged texts were less accurately predictable than the positive ones. This may be caused because negative opinions involve sarcasm, irony and such complicated semantics that are difficult to be detected by machines. Undoubtedly, there is a lot of place for further improvement. Developing novel soft computing methodologies to model and predict language structure of OSN so as to be fed as parameters in machine learning algorithms is one promising idea. Another possible optimization could be reached by incorporating graph theory and features like author's position in the network. Although interestingly are mixed the Semantic methods such as entity extraction with the statistical methods of machine learning to perform the SA task, they focus on static characteristics of texts, neglecting the fact that each message has a networked context. Incorporating network structures and relations in text analysis gives a more systemic view, which achieves more reliable result on text analysis. The unique element of social networked data is after all, that they reveal information about interactions between users-communities-content.

11 CONCLUSIONS AND FUTURE RESEARCH

The emerging paradigm of social networking and big data analytics provides enormous research challenges. In this research, we review studies on applying data analysis to the social networked data, which consists of social network analysis and content based analysis, from the perspective of techniques and frameworks. Since, the efficiency of every information processing procedure is greatly affected by the quality of the data, data cleansing frameworks are also investigated.

The long standing statistical issues of representativeness, biases and data-cleaning subjectivity are applied even greater to social networks. In the case of social networks, the extraction of data is highly restricted by platform's terms of service, API transactions and usability. Among all the social networks the Twitter is by far the most analyzed platform due to its API flexibility. However, a research area that depends on a single data source, as compelling as it is, entails many perils. A distinct evidence is that, resources and results from previous researches are based on twitter vocabulary which force future studies to utilize Twitter again, hence perpetuating a vicious cycle. Monitoring and analyzing events, activities and sentiments from different social media services remains a challenge. Also utilizing cross collaborative recommendation where information across multiple recommender systems is a new frontier. On the one hand this would be quite reasonable since social networks contains data mostly about people and privacy preserving is yet to be accomplished. On the other hand, in the rapidly-evolving data economy such data has become the new currency where only governments and enterprises have the privilege to explore them, provoking in this way the "Data Democracy" struggling.

Considering the data analysis development environment, near real time analysis via online algorithms scalable in memory and computational resources, is required. Cloud-oriented processing techniques can meet computational needs and the performance required in fast extraction of data from social networking sites. They are widely utilized in data analytics but, to the best of our knowledge, there isn't yet any application on data cleaning.

The collected data are hard to be analyzed because of their unique discourse structure and

grammar, at which conventional preprocessing tools are susceptible. Most researchers have faced enormous difficulties in dealing with the noise of OSN's. Resources that could alleviate this issue (e.g. a publicly available large corpus of posted messages in order to find patterns in informal discourse language) is not easy to be built mostly due to privacy concerns.

Arguably, processing text at word level is not a reliable option. In general, current NLP methods are considered insufficient because they mostly focus on word co-occurrences frequencies neglecting the complex nature of human language which is neither a set of mere words nor a machine understandable one. Challenges such as the cascade of the semantically related concepts that every word unfolds, irony, sarcasm, previously unseen words, word ambiguities and syntactic complexities are some indicators showing that the problem of interpreting human language cannot be translated into binary language for computers to process it, and a deep understanding of natural language by machines is needed.

After an efficient data preprocessing without discarding useful information, typically follows the data representation step which should go from bag of words to bag of concepts and even better to bag of narratives level. Sentic computing and semantic technologies provide some possible solutions to these problems, but they cannot fully solve them because these two technologies are not mature enough.

Apart from concept level analysis, equal promising is Computational Intelligence paradigm. Its inherent ability of collective intelligence and adaptability to a changing environment are significant features that can help in social network data analysis. Deep learning has been preferred in the recent studies responding to the need of how machines could "understand" the text instead of mere "see" it. However, more robust solutions would be provided by integrating many social network sources.

For all these reasons, it is quite reasonable why social network data collection, scrubbing and analysis still demands a remarkable collection of tools and skills. In this end, extracting insights from social networked data are still far from perfect. This is an area of research that emerged during the last years, in parallel with the growth of user participation in social networks. After reporting on the most recent efforts in the area, it

is clear that there is a lot of space for improvement towards this direction. To help the research community and the audience of the research to find *something* to proceed, the broader issue is broken into possible high impact research trends for future work:

- Reconciling data, ensuring consistency across sources, checking the quality of data seems to be the hard part of big data; and validating the accuracy of machine-generated data without the assessment of human intervention would be definitely a great breakthrough.
- Cleaning and preprocessing data of OSN in cloud.
- Early event detection by integrating data from different OSN. Since social networks connect people who expose similar interests, the patterns in the content they share or the relationships they form are differentiated across OSN. Though, this presupposes that other OSN, apart from Twitter, give access to their data through API's. In this end, a great challenge that has to be addressed is the lack of flexible API's.
- The bag of words (BoW) model should be replaced with more sophisticated data representation of bag of concept model. Taking into account, semantic relations between words can predict discourse structures such as comparison which can result in more accurate sentiment analysis.
- Developing resources towards the data scope of social networks that can handle informal text better is indeed a current need. Available resources are normally trained on corpora of full text documents such as news wire articles, which are very different from tweets in terms of length and content. For instance, dependency parsers, like the Stanford Parser, doesn't handle ungrammatical text very well because they were trained on Wall Street Journal.
- Sentiment analysis calculated via an ordered five-point scale metrics is under investigation together with the quantification of *prevalence* of positive and negative tweets about a given topic.
- An opinion tagger/classifier that detects opinionated text and no-opinionated one isn't closely studied.
- An emotion classifier, a comparative opinion identifier and a spam opinion tracker are all interesting topics.
- Improved filtering, detection and analysis algorithms that utilize Computational Intelligence, deep learning and discourse relations on information from multiple

sources and multiple languages is another challenging area.

- Context-aware systems are also hot research areas. Context-aware methods identify ambiguous terms that vary in meaning depending on the context they are expressed.
- A sub-field of sentiment analysis that is becoming increasingly popular is multimodal sentiment analysis.
- The diffusion of sentiments in the social network and people's sentiments correlation between internal (their friends) and external (public events) factors is also an option for research.
- Corporation of methods from complex network systems in understanding individual behaviors and collective cognition in social networks is a new scientific interest.
- In the field of TDT, an interest is shaped towards discovering efficient methods in detecting fake reviews.
- Deep learning and Computing Intelligence have shown potential as the basis for software that could work out the emotions or events described in text even if they aren't explicitly referenced. They can also recognize objects in photos, and make sophisticated predictions about people's likely future behavior. Yet, data analysis applications and tools in big data social networking utilizing such promising techniques are missing. Ranging from comparing methods' efficiency on specific analysis tasks to building frameworks, are all hype research topics.
- Privacy- preserving collaborative filtering (PPCF) in social recommender systems is a recent challenging topic.
- How the determination of data trustworthiness, the identification of errors and how the biases are evaluated and corrected is not yet explored in OSN.

A considerable effort is still required to achieve efficient and reliable analysis systems that exploit this rich and continuous flow of user-generated content and social relations. It is expected that as social networks sources emerge, social network analysis and content mining will remain significant and challenging. Generally, investments on how social data are collected and cleaned should be done so as the accuracy of the analytical results to be proven and automated tools to be built.

12 REFERENCES

BOOKS:

C. C. Aggarwal, Ed., *Social Network Data Analytics*. Boston, MA: Springer US, 2011.

C. D. Manning, P. Raghavan, and H. Schütze, "Scoring, term weighting and the vector space model," in *Introduction to Information Retrieval*, Cambridge University Press, 2009.

C. Shuguang, O. H. Alfred, L. Zhi-Quan, and M. F. M. José, *Big Data Over Networks*. 2016.

E. Cambria and A. Hussain, "Introduction," in *Sentic Computing*, Cham: Springer International Publishing, 2015, pp. 1–21.

E. D. Kolaczyk and G. Csárdi, *Statistical Analysis of Network Data with R*. Springer, 2014.

E. David and K. Jon, "Networks, Crowds, and Markets: Reasoning About a Highly Connected World," Jul. 2010.

E. R. Babbie, (2007). "The practice of social research. Belmont, CA.: Thomson Wadsworth".

F. Frasincar, W. IJntema, F. Goossen, F. Hogenboom, G. Adomavicius, A. Tuzhilin, D. W. Aha, D. Kibler, M. K. Albert, C. Buckley, J. Allan, G. Salton, C. C. Chen, M. C. Chen, F. Frasincar, J. Borsje, L. Levering, P. Jaccard, H. P. Luhn, S. E. Middleton, N. R. Shadbolt, D. C. De Roure, B. Pang, L. Lee, G. Salton, C. Buckley, A. Singhal, G. Salton, M. Mitra, and C. Buckley, "A Semantic Approach for News Recommendation," in *Business Intelligence Applications and the Web*, vol. 17, no. 6, IGI Global, 1AD, pp. 102–121.

Hansen D., Shneiderman B., and M. A. Smith, *Analyzing Social Media Networks with NodeXL: Insights from a Connected World*. Morgan Kaufmann, 2010.

K. Cherven, *Network Graph Analysis and Visualization with Gephi*. Packt Publishing Ltd, 2013.

K. Pattnaik, B.S. Prasad Mishra, Introduction to Big Data Analysis, in:*Techniques and Environments for Big Data Analysis* 2016: pp. 1–20.

M. A. H. Al-Hamami, D. Agrawal, S. Das, A. El Abbadi, S. Barnes, D. Boyd, K. Crawford, H. Cheng, C. Chew, G. Eysenbach, J. C. Chuang, H. Y. Chen, T. Colin, C. Coronel, S. Morris, P. Rob, N. Daniel, M. Di Domenico, M. Dodge, Z. Feng, X. Hui-Feng, X. Dung-Sheng, Z. Yong-Heng, Y. Fei, L. Manovich, H. F. Qin, Z. H. Li, H. Wang, and T. Yang, "The Impact of Big Data on Security," in *Handbook of Research on Threat Detection and Countermeasures in Network Security*, vol. 3, IGI Global, 2015, pp. 276–298.

M. Huisman, M.A. Van Duijn, Software for Social Network Analysis, in: Model. Methods Soc. Netw. Anal., Cambridge University Press, 2005: pp. 270–316. https://books.google.com/books?hl=en&lr=&id=4Ty5xP_KcpAC&pgis=1 (accessed April 21, 2016)

M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, E. Muharemagic, "Deep learning applications and challenges in big data analytics," *J. Big Data*, vol. 2, no. 1, p. 1, Dec. 2015.

M. Vaidya, "Handling Critical Issues of Big Data on Cloud," in *Managing Big Data in Cloud Computing Environments*, vol. 1, IGI Global, 2016, pp. 100–131.

N. Siddique and H. Adeli, "Introduction to Computational Intelligence," in *Computational Intelligence: Synergies of Fuzzy Logic, Neural Networks and Evolutionary Computing*, Oxford, UK: John Wiley & Sons Ltd, 2013, pp. 1–17.

P. Chaudhary, S. Gupta, B. B. Gupta, V. S. Chandra, S. Selvakumar, M. Fire, R. Goldschmidt, Y. Elovici, S. Gangwar, M. Kumar, P. K. Meena, S. Gupta, and L. Sharma, "Auditing Defense against XSS Worms in Online Social Network-Based Web Applications," in *Handbook of Research on Modern Cryptographic Solutions for Computer and Cyber Security*, vol. 36, no. 5, IGI Global, 1AD, pp. 216–245.

P.J. Sadalage, M.Fowler, "NoSQL Distilled A Brief Guide to the Emerging World of Polyglot Persistence"2012

R. A. Hanneman and M. Riddle, *Introduction to Social Network Methods*. 2005.

R. Alhajj and J. Rokne, Eds., *Encyclopedia of Social Network Analysis and Mining*. New York, NY: Springer New York, 2014.

Raj, Pethuru, The Compute Infrastructures for Big Data Analytics., in: Handb. Res. Cloud Infrastructures Big Data Anal. (Advances Data Min. Database Manag. B. Ser. (978, Information Science Reference, 2014. https://www.amazon.com/Handbook-Research-Infrastructures-Analytics-Management/dp/1466658649

R. K. Yin, (2009). "Case study research: design and methods". Los Angeles [etc.]: Sage.

S. Mithun, "Exploiting Rhetorical Relations in Blog Summarization," in *Advances in Artificial Intelligence*, Springer Berlin Heidelberg, 2010, pp. 388–392

S. Prom-on, S. N. Ranong, P. Jenviriyakul, T. Wongkaew, N. Saetiew, and T. Achalakul, "DOM: A big data analytics framework for mining Thai public opinions," in *Big Data: Principles and Paradigms*, R. Buyya, Ed. Morgan Kaufmann, 2016, pp. 339–355.

S. Bandyopadhyay, A.R. Rao, B.K. Sinha, Introduction to Social Network Analysis, in: Model. Soc. Networks with Stat. Appl., SAGE, 2011: p. 235

T. Hey, The Fourth Paradigm – Data-Intensive Scientific Discovery, Springer Berlin Heidelberg, 2012. doi:10.1007/978-3-642-33299-9_1.

V. Bhatnagar, "Data Mining in Dynamic Social Networks and Fuzzy Systems,", IGI Global, 2013.

W. Chang, J. Wu, L. M. Aiello, A. Barrat, R. Schifanella, C. Cattuto, B. Markines, F. Menczer, A. Blum, K. Ligett, A. Roth, C. Dwork, S. Kisilevich, L. Rokach, Y. Elovici, B. Shapira, Y. Li, M. Chen, Q. Li, W. Zhang, J. Lin, A. Machanavajjhala, D. Kifer, J. Gehrke, M. Venkitasubramaniam, P. Samarati, L. Sweeney, and L. Sweeney, "A New View of Privacy in Social Networks:," in *Handbook of Research on Modern Cryptographic Solutions for Computer and Cyber Security*, vol. 6, no. 2, IGI Global, 1AD, pp. 28–51,2016.

JOURNALS

Á. Cuesta, D. F. . Barrero, and M. D. R-Moreno, "A FRAMEWORK FOR MASSIVE TWITTER DATA EXTRACTION AND ANALYSIS," *Malaysian J. Comput. Sci.*, vol. 27, no. 1, p. 50, 2014.

A. Abaid, A. Elmagarmid, I. Ilyas, M. Ouzzani, J.-A. Quiane-Ruiz, N. Tang, S. Yin, A. Ebaid, I.F. Ilyas, Purdue e-Pubs NADEEF: A Generalized Data Cleaning System NADEEF: A Generalized Data Cleaning System, (2013). http://docs.lib.purdue.edu/ccpubs.

Abbas, L. Zhang, and S. U. Khan, "A survey on context-aware recommender systems based on computational intelligence techniques," *Computing*, vol. 97, no. 7, pp. 667–690, Jul. 2015.

Atefeh and W. Khreich, "A Survey of Techniques for Event Detection in Twitter," *Comput. Intell.*, vol. 31, no. 1, pp. 132–164, Feb. 2015.

B. Perozzi, R. Al-Rfou, and S. Skiena, "DeepWalk," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*, 2014, pp. 701–710.

Batrinca and P. C. Treleaven, "Social media analytics: a survey of techniques, tools and platforms" *AI Soc.*, vol. 30, no. 1, pp. 89–116, Feb. 2015.

Boyd D. and K. Crawford, "CRITICAL QUESTIONS FOR BIG DATA," *Information, Commun. Soc.*, vol. 15, no. 5, pp. 662–679, Jun. 2012.

C. Musto, G. Semeraro, and M. Polignano, "A comparison of Lexicon-based approaches for Sentiment Analysis of microblog posts," in *Proceedings of the 8th International Workshop on Information Filtering and Retrieval Workshop of the XIII AI*IA Symposium on Artificial Intelligence*, 2014, pp. 59–68.

C.-C. Hsu, H.-C. Chen, K.-K. Huang, and Y.-M. Huang, “A personalized auxiliary material recommendation system based on learning style on Facebook applying an artificial bee colony algorithm,” *Comput. Math. with Appl.*, vol. 64, no. 5, pp. 1506–1513, 2012.

C.-W. Tsai, C.-F. Lai, H.-C. Chao, and A. V. Vasilakos, “Big data analytics: a survey,” *J. Big Data*, vol. 2, no. 1, p. 21, Oct. 2015.

Cagliero and A. Fiori, “TweCoM: Topic and Context Mining from Twitter,” Springer Vienna, 2013, pp. 75–100.

Cai and Y. Zhu, “The Challenges of Data Quality and Data Quality Assessment in the Big Data Era,” *Data Sci. J.*, vol. 14, no. 0, p. 2, May 2015.

C. Fürber, M. Hepp, Towards a vocabulary for data quality management in semantic web architectures, in: Proc. 1st Int. Work. Linked Web Data Manag. - LWDM ’11, ACM Press, New York, New York, USA, 2011: p. 1. doi:10.1145/1966901.1966903.

D. B. Kurka, A. Godoy, and F. J. Von Zuben, “Online Social Network Analysis: A Survey of Research Applications in Computer Science,” Apr. 2015.

D.-H. Shin, M.J. Choi, Ecological views of big data: Perspectives and issues, Telemat. Informatics. 32 (2015) 311–320. doi:10.1016/j.tele.2014.09.006.

D. Kotzias, M. Denil, N. de Freitas, P. Smyth, From Group to Individual Labels Using Deep Features, in: Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD ’15, ACM Press, New York, New York, USA, 2015: pp. 597–606. doi:10.1145/2783258.2783380.

H. Mackay, Information and the transformation of sociology: interactivity and social media monitoring, Commun. Capital. Crit. 11 (2013) 117–126.

H. Saif, M. Fernández, Y. He, H. Alani, Evaluation datasets for Twitter sentiment analysis: a survey and a new dataset, the STS-Gold, in: 1st Interantional Work. Emot. Sentim. Soc. Expressive Media Approaches Perspect. from AI, 2013.

D. T. Nguyen and J. E. Jung, “Real-time event detection for online behavioral analysis of big social data,” *Futur. Gener. Comput. Syst.*, 2016.

E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, “The role of social networks in information diffusion,” in *Proceedings of the 21st international conference on World Wide Web - WWW ’12*, 2012, p. 519.

E. Cambria, A. Hussain, C. Havasi, and C. Eckl, “Sentic Computing: Exploitation of Common Sense for the Development of Emotion-Sensitive Systems,” in *Proceedings of the Second international conference on Development of Multimodal Interfaces: active Listening and Synchrony*, Springer-Verlag, 2010, pp. 148–156.

E. Enhua Tan, L. Lei Guo, S. Songqing Chen, X. Xiaodong Zhang, Y. Yihong Zhao, Spammer Behavior Analysis and Detection in User Generated Content on Social Networks, in: 2012 IEEE 32nd Int. Conf. Distrib. Comput. Syst., IEEE, 2012: pp. 305–314. doi:10.1109/ICDCS.2012.40.

E. Cambria, B. White, Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article], IEEE Comput. Intell. Mag. 9 (2014) 48–57. doi:10.1109/MCI.2014.2307227

E. Davis and G. Marcus, “Commonsense reasoning and commonsense knowledge in artificial intelligence,” *Commun. ACM*, vol. 58, no. 9, pp. 92–103, Aug. 2015.

E. J. Newman, “The Structure and Function of Complex Networks,” *SIAM Rev.*, vol. 45, no. 2, pp. 167–256, Jan. 2003.

E. Kontopoulos, C. Berberidis, T. Dergiades, and N. Bassiliades, “Ontology-based sentiment analysis of twitter posts,” *Expert Syst. Appl.*, vol. 40, no. 10, pp. 4065–4074, 2013.

E. Rahm and H. H. Do, “Data Cleaning: Problems and Current Approaches, Microsoft Research, 2000”

Enhua Tan, Lei Guo, Songqing Chen, Xiaodong Zhang, and Yihong Zhao, “Spammer

Behavior Analysis and Detection in User Generated Content on Social Networks," in *2012 IEEE 32nd International Conference on Distributed Computing Systems*, 2012, pp. 305–314.

F. Bonchi, C. Castillo, A. Gionis, and A. Jaimes, "Social Network Analysis and Mining for Business Applications," *ACM Trans. Intell. Syst. Technol. Artic.*, vol. 2, no. 22, 2011.

F. Geerts, G. Mecca, P. Papotti, and D. Santoro, "The LLUNATIC Data-Cleaning Framework," *VLDB Endow.*, vol. 6, no. 9, 2013.

F. L. F. Almeida and C. Calistru, "The main challenges and issues of big data management," *Int. J. Res. Stud. Comput.*, vol. 2, no. 1, Apr. 2013.

Fire, L. Tenenboim-Chekina, R. Puzis, O. Lesser, L. Rokach, and Y. Elovici, "Computationally efficient link prediction in a variety of social networks," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 1, pp. 1–25, Dec. 2013.

Fu, Z. Niu, C. Zhang, J. Ma, and J. Chen, "Visual Cortex Inspired CNN Model for Feature Construction in Text Analysis," *Front. Comput. Neurosci.*, vol. 10, p. 64, Jul. 2016.

G. Anthes, "Data brokers are watching you," *Commun. ACM*, vol. 58, no. 1, pp. 28–30, Dec. 2014.

Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *Int. J. Inf. Manage.*, vol. 35, no. 2, pp. 137–144, Apr. 2015.

H. Agrawal, A. Thakur, and R. Slathia, "A Comparative Analysis of Social Networking Analysis Tools," *J. Inf. Technol. Softw. Eng.*, vol. 05, no. 03, Oct. 2015.

H. Chen, X. Cui, and H. Jin, "Top-k followee recommendation over microblogging systems by exploiting diverse information sources," *Futur. Gener. Comput. Syst.*, vol. 55, pp. 534–543, 2016.

H. V. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, and C. Shahabi, "Big data and its technical challenges," *Commun. ACM*, vol. 57, no. 7, pp. 86–94, Jul. 2014.

Hassan and A. Menezes, "Social Text Normalization using Contextual Graph Random Walks," pp. 1577–1586, 2013.

Hu, Y. Wen, T. S. Chua, and X. Li, "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial," *IEEE Access*, vol. 2, pp. 652–687, 2014.

I. Maletic and A. Marcus, "Data Cleansing: A Prelude to Knowledge Discovery," in *Data Mining and Knowledge Discovery Handbook*, Boston, MA: Springer US, 2009, pp. 19–32.

Immonen, Pääkkönen P., Ovaska E."Evaluating the Quality of Social Media Data in Big Data Architecture",IEEE Access, vol 3, Oct. 2015.

J. He and W. W. Chu, "A Social Network-Based Recommender System (SNRS)," Springer US, 2010, pp. 47–74

J. L. Reyes-Ortiz, L. Oneto, and D. Anguita, "Big Data Analytics in the Cloud: Spark on Hadoop vs MPI/OpenMP on Beowulf," *Procedia Comput. Sci.*, vol. 53, pp. 121–130, 2015.

J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling, "TwitterStand," in *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '09*, 2009, p. 42.

J. Wang, S. Krishnan, M. J. Franklin, K. Goldberg, T. Kraska, and T. Milo, "A sample-and-clean framework for fast and accurate query processing on dirty data," in *Proceedings of the 2014 ACM SIGMOD international conference on Management of data - SIGMOD '14*, 2014, pp. 469–480.

K. Cukier and V. Mayer-Schoenberger, "The Rise of Big Data: How It's Changing the Way We Think About the World," *FOREIGN Aff.*, vol. 92, no. 3, pp. 28–40, 2013.

K. Natarajan, J. Li, and A. Koronios, “Data mining techniques for data cleaning,” in *Engineering Asset Lifecycle Management*, London: Springer London, 2010, pp. 796–804

K. Schouten and F. Frasincar, “Survey on Aspect-Level Sentiment Analysis,” *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 3, pp. 813–830, Mar. 2016.

Khalid, M. U. S. Khan, S. U. Khan, and A. Y. Zomaya, “OmniSuggest: A Ubiquitous Cloud-Based Context-Aware Recommendation System for Mobile Social Networks,” *IEEE Trans. Serv. Comput.*, vol. 7, no. 3, 2014.

G. Kossinets and D. J. Watts, “Empirical Analysis of an Evolving Social Network,” *Science (80-.).*, vol. 311, no. 5757, pp. 88–90, 2006.

L. Staudt, A. Sazonovs, and H. Meyerhenke, “NetworKit: A Tool Suite for Large-scale Complex Network Analysis,” p. 21, Mar. 2014.

Labrinidis and H. V. Jagadish, “Challenges and opportunities with big data,” *Proc. VLDB Endow.*, vol. 5, no. 12, pp. 2032–2033, Aug. 2012

Lee, Y. W., D. M. Strong, B. K. Kahn and R. Y. Wang (2002). “AIMQ: A Methodology for Information Quality Assessment.”Information & Management40 (2), 133-146.

Li, C. Chen, Q. Lv, L. Shang, Y. Zhao, T. Lu, and N. Gu, “An algorithm for efficient privacy-preserving item-based collaborative filtering,” *Futur. Gener. Comput. Syst.*, vol. 55, pp. 311–320, 2016.

Liu and L. Zhang, “A Survey of Opinion Mining and Sentiment Analysis,” in *Mining Text Data*, Boston, MA: Springer US, 2012, pp. 415–463.

L. Kolb, A. Thor, E. Rahm, Dedoop: Efficient Deduplication with Hadoop, in: Proc. VLDB Endow., 2012.

Lj. Sheela, “A Review of Sentiment Analysis in Twitter Data Using Hadoop,” *Int. J. Database Theory Appl.*, vol. 9, no. 1, pp. 77–86, 2016.

M. A. Smith, “NodeXL: Simple network analysis for social media,” in *2013 International Conference on Collaboration Technologies and Systems (CTS)*, 2013, pp. 89–93.

M. A. Smith, B. Shneiderman, N. Milic-Frayling, E. Mendes Rodrigues, V. Barash, C. Dunne, T. Capone, A. Perer, and E. Gleave, “Analyzing (social media) networks with NodeXL,” in *Proceedings of the fourth international conference on Communities and technologies - C&T ’09*, 2009, p. 255.

M. Adedoyin-Olowe, M. M. Gaber, and F. Stahl, “A Survey of Data Mining Techniques for Social Network Analysis,” *J. Data Min. Digit. Humanit.*, 2014.

M. Chen, S. Mao, and Y. Liu, “Big Data: A Survey,” *Mob. Networks Appl.*, vol. 19, no. 2, pp. 171–209, Jan. 2014.

M. Fire, L. Tenenboim-Chekina, R. Puzis, O. Lesser, L. Rokach, and Y. Elovici, “Computationally efficient link prediction in a variety of social networks,” *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 1, pp. 1–25, Dec. 2013.

M. Mathioudakis and N. Koudas, “TwitterMonitor,” in *Proceedings of the 2010 international conference on Management of data - SIGMOD ’10*, 2010, p. 1155.

M. Mezzanzanica, R. Boselli, M. Cesarini, and F. Mercorio, “Improving Data Cleansing Accuracy: A model-based Approach.”, 2014

M. Speriosu, N. Sudan, S. Upadhyay, J. Baldridge, Twitter Polarity Classification with Label Propagation over Lexical Links and the Follower Graph, in: Proc. EMNLP 2011, Conf. Empir. Methods Nat. Lang. Process., 2011: pp. 53–63.

M. Tanwar, R. Duggal, and S. K. Khatri, “Unravelling unstructured data: A wealth of information in big data,” in *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)*, 2015, pp. 1–6.

M. Volkovs, F. Chiang, J. Szlichta, and R. J. Miller, "Continuous Data Cleaning.", *30th IEEE International Conference on Data Engineering,* 2014

Maiya, A. S., & Berger-Wolf, T. Y. (2011). Benefits of bias: Towards better characterization of network sampling. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Meng X. and Ci X., "Big Data Management: Concepts,Techniques and Challenges"[J]. *计算机研究与发展,* 2013, 50(1): 146-169.

Minanovic, H. Gabelica, and Z. Krstic, "Big data and sentiment analysis using KNIME: Online reviews vs. social media," in *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2014, pp. 1464–1468.

Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel, "You are who you know," in *Proceedings of the third ACM international conference on Web search and data mining - WSDM '10*, 2010, p. 251.

Mrvar and V. Batagelj, "Analysis and visualization of large networks with program package Pajek," *Complex Adapt. Syst. Model.*, vol. 4, no. 1, p. 6, Apr. 2016.

M. S. Handcock, D. R. Hunter, C. T. Butts, S. M. Goodreau, and M. Morris, "statnet: Software Tools for the Representation, Visualization, Analysis and Simulation of Network Data.," *J. Stat. Softw.*, vol. 24, no. 1, pp. 1548–7660, 2008.

N. Akhtar, "Social Network Analysis Tools," in *2014 Fourth International Conference on Communication Systems and Network Technologies*, 2014, pp. 388–392

N. Khan, I. Yaqoob, I. A. T. Hashem, Z. Inayat, M. Mahmoud Ali, Waleed Kamaleldin Alam, M. Shiraz, and A. Gani, "Big Data: Survey, Technologies, Opportunities, and Challenges," *Sci. World J.*, vol. 2014, 2014.

N. Panagiotou, I. Katakis, and D. Gunopulos, *Detecting Events in Online Social Networks: Definitions, Trends and Challenges*. 2016.

N. Tang, "Big Data Cleaning," Springer International Publishing, 2014, pp. 13–24.

Peled, M. Fire, L. Rokach, and Y. Elovici, "Matching Entities Across Online Social Networks," Oct. 2014.

Q. Wang, J. She, T. Song, Y. Tong, L. Chen, and K. Xu, "Adjustable Time-Window-Based Event Detection on Twitter," Springer International Publishing, 2016, pp. 265–278.

R. Boselli, M. Cesarini, F. Mercorio, and M. Mezzanzanica, "Accurate Data Cleansing through Model Checking and Machine Learning Techniques," Springer International Publishing, 2015, pp. 62–80.

R. Kaushik, S. Apoorva Chandra, D. Mallya, J. N. V. K. Chaitanya, and S. S. Kamath, "Sociopedia: An Interactive System for Event Detection and Trend Analysis for Twitter Data," Springer India, 2016, pp. 63–70.

R. Mihalcea, C. Banea, J. Wiebe, *"Learning Multilingual Subjective Language via Cross-Lingual Projections"*http://www.aclweb.org/anthology/P07-1123 (PDF). Proceedings of the Association for Computational Linguistics (ACL). pp. 976–983, 2007.

R. Moraes, J. F. Valiati, and W. P. Gavião Neto, "Document-level sentiment classification: An empirical comparison between SVM and ANN," *Expert Syst. Appl.*, vol. 40, no. 2, pp. 621–633, 2013.

Ritter, Mausam, O. Etzioni, and S. Clark, "Open domain event extraction from twitter," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12*, 2012, p. 1104.

S. A. Myers, A. Sharma, P. Gupta, and J. Lin, "Information network or social network?," in *Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion*,

2014, pp. 493–498.

S. Gill, B. Lee, and E. Neto, "Context aware model-based cleaning of data streams" in *2015 26th Irish Signals and Systems Conference (ISSC)*, 2015.

S. Heymann and B. Le Grand, "Visual Analysis of Complex Networks for Business Intelligence with Gephi," in *2013 17th International Conference on Information Visualisation*, 2013, pp. 307–312.

S. Kandel, A. Paepcke, J. Hellerstein, J. Heer, Wrangler, in: Proc. 2011 Annu. Conf. Hum. Factors Comput. Syst. - CHI '11, ACM Press, New York, New York, USA, 2011: p. 3363. doi:10.1145/1978942.1979444.

S. Kaisler, F. Armour, J.A. Espinosa, W. Money, Big Data: Issues and Challenges Moving Forward, in: 2013 46th Hawaii Int. Conf. Syst. Sci., IEEE, 2013: pp. 995–1004. doi:10.1109/HICSS.2013.645.

S. Poria, E. Cambria, G. Winterstein, and G.-B. Huang, "Sentic patterns: Dependency-based rules for concept-level sentiment analysis," *Knowledge-Based Syst.*, vol. 69, pp. 45–63, 2014.

S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," *Neurocomputing*, vol. 174, pp. 50–59, 2016.

S. Rill, D. Reinel, J. Scheidt, and R. V. Zicari, "PoliTwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis," *Knowledge-Based Syst.*, vol. 69, pp. 24–33, 2014.

S. Sulaiman, S. A. Yahya, N. S. M. Shukor, A. R. Ismail, Q. Zaahirah, H. Yaacob, A. W. A. Rahman, and M. A. Dzulkifli, "Clustering Natural Language Morphemes from EEG Signals Using the Artificial Bee Colony Algorithm," Springer International Publishing, 2015, pp. 51–60.

S.P. Ahuja and B. Moore, "State of Big Data Analysis in the Cloud," *Netw. Commun. Technol.*, vol. 2, no. 1, p. 62, May 2013.

Saha and D. Srivastava, "Data quality: The other face of Big Data," in *2014 IEEE 30th International Conference on Data Engineering*, 2014, pp. 1294–1297.

Schmidt, M. Atzmueller, M. Hollender, "Data Preparation for Big Data Analytics:," in *Enterprise Big Data Engineering, Analytics, and Management*, vol. 5, IGI Global, 2016, pp. 157–170.

Shah and T. Zaman, "Community Detection in Networks: The Leader-Follower Algorithm," Nov. 2010.

Sieg, B. Mobasher, and R. Burke, "Improving the effectiveness of collaborative recommendation with ontology-based user profiles," in *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems - HetRec '10*, 2010, pp. 39–46.

T. Hashem, I. Yaqoob, N. Badrul Anuar, S. Mokhtar, A. Gani, and S. Ullah Khan, "The rise of 'Big Data' on cloud computing: Review and open research issues," *Inf. Syst.*, vol. 47, pp. 98–115, Aug. 2014.

T. MA, J. ZHOU, M. TANG, Y. TIAN, A. AL-DHELAAN, M. AL-RODHAAN, and S. LEE, "Social Network and Tag Sources Based Augmenting Collaborative Recommender System," *IEICE Trans. Inf. Syst.*, vol. E98-D, no. 4, pp. 902–910, 2015.

T. Rambharose and A. Nikov, "Computational intelligence-based personalization of interactive web systems," *WSEAS Trans. Inf. Sci. Appl.*, vol. 7, no. 4, pp. 484–497, 2010.

T. Zhu, Y. Ren, W. Zhou, J. Rong, and P. Xiong, "An effective privacy preserving algorithm for neighborhood-based collaborative filtering," *Futur. Gener. Comput. Syst.*, vol. 36, pp. 142–155, 2014.

Taleb, R. Dssouli, and M. A. Serhani, "Big Data Pre-processing: A Quality Framework," in *2015 IEEE International Congress on Big Data*, 2015, pp. 191–198.

V. C. M. Leung, "Enabling technologies for future data center networking: a primer," *IEEE Netw.*, vol. 27, no. 4, pp. 8–15, 2013.

V. S. V. Pulla, C. Varol, and M. Al, "Open Source Data Quality Tools: Revisited," Springer International Publishing, 2016, pp. 893–902.

V.J. Nirmal, D.I.G. Amalarethinam, C. Author, Parallel Implementation of Big Data Pre-Processing Algorithms for Sentiment Analysis of Social Networking Data, Intern. J. Fuzzy Math. Arch. 6 (2015) 149–159. www.researchmathsci.org

Vakali, M. Giatsoglou, and S. Antaris, "Social networking trends and dynamics detection via a cloud-based framework design," in *Proceedings of the 21st international conference companion on World Wide Web - WWW '12 Companion*, 2012, p. 1213.

W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, 2014.

W. Shi, Y. Zhu, T. Huang, G. Sheng, Y. Lian, G. Wang, and Y. Chen, "An Integrated Data Preprocessing Framework Based on Apache Spark for Fault Diagnosis of Power Grid Equipment," *J. Signal Process. Syst.*, pp. 1–16, Mar. 2016.

W. Tan, M. B. Blake, I. Saleh, and S. Dustdar, "Social-Network-Sourced Big Data Analytics," *IEEE Internet Comput.*, vol. 17, no. 5, pp. 62–69, Sep. 2013.

W. Wambeke, M. Liu, and S. M. Hsiang, "Using Pajek and Centrality Analysis to Identify a Social Network of Construction Trades," *J. Constr. Eng. Manag.*, vol. 138, no. 10, pp. 1192–1201, Oct. 2012

Wolfswinkel, J. F., Furtmueller, E., & Wilderom, C. P. M. (2013). Using grounded theory as a

Wu, D. Schaefer, and D. W. Rosen, "Cloud-based design and manufacturing systems: A social network analysis," in *ICED13: 19th International Conference on Engineering Design*, 2013

X. Chu, J. Morcos, I. F. Ilyas, M. Ouzzani, P. Papotti, N. Tang, and Y. Ye, "KATARA," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data - SIGMOD '15*, 2015, pp. 1247–1261.

X. Chu, I.F. Ilyas, S. Krishnan, J. Wang, Data Cleaning: Overview and Emerging Challenges, in: SIGMOD '16 Proc. 2016 Int. Conf. Manag. Data, 2016: pp. 2201–2206. doi:10.1145/2882903.2912574.

X. Chu"Holistic Data Cleaning: Putting Violations Into Context.", *in Data Enginnering IEEE*, 2013

X. Zheng, Z. Zeng, Z. Chen, Y. Yu, C. Rong, Detecting spammers on social networks, Neurocomputing. 159 (2015) 27–34. doi:10.1016/j.neucom.2015.02.047.

X. Hu, L. Tang, J. Tang, and H. Liu, "Exploiting social relations for sentiment analysis in microblogging," in *Proceedings of the sixth ACM international conference on Web search and data mining - WSDM '13*, 2013, p. 537.

X. Yang, B. Pan, J. A. Evans, and B. Lv, "Forecasting Chinese tourist volume with search engine data," *Tour. Manag.*, vol. 46, pp. 386–397, 2015.

X. Yang, Y. Guo, Y. Liu, and H. Steck, "A survey of collaborative filtering based social recommender systems," *Comput. Commun.*, vol. 41, pp. 1–10, 2014.

X. Zou, W. Zhu, L. Yang, and Y. Shu, "[Google Flu Trends--the initial application of big data in public health].," *Zhonghua Yu Fang Yi Xue Za Zhi*, vol. 49, no. 6, pp. 581–4, Jun. 2015.

Y. Cao, W. Fan, W. Yu, Determining the Relative Accuracy of Attributes, in: Proc. 2013 ACM SIGMOD Int. Conf. Manag. Data, ACM New York, 2013: pp. 565–576.

Y. Yu, H. Lin, J. Meng, and Z. Zhao, "Visual and Textual Sentiment Analysis of a Microblog Using Deep Convolutional Neural Networks," *Algorithms*, vol. 9, no. 2, p. 41, Jun. 2016.

Y.-M. Li and T.-Y. Li, "Deriving market intelligence from microblogs," *Decis. Support Syst.*, vol. 55, no. 1, pp. 206–217, 2013.

Z. Abedjan, C.G. Akcora, M. Ouzzani, P. Papotti, M. Stonebraker, Temporal rules discovery for web data cleaning, Proc. VLDB Endow. 9 (2015) 336–347. doi:10.14778/2856318.2856328.

Z. Fu, K. Ren, J. Shu, X. Sun, and F. Huang, "Enabling Personalized Search over Encrypted Outsourced Data with Efficiency Improvement," *IEEE Trans. Parallel Distrib. Syst.*, vol. 27, no. 9, pp. 2546–2559, Sep. 2016

Z. FU, X. SUN, Q. LIU, L. ZHOU, and J. SHU, "Achieving Efficient Cloud Search Services: Multi-Keyword Ranked Search over Encrypted Cloud Data Supporting Parallel Computing," *IEICE Trans. Commun.*, vol. E98-B, no. 1, pp. 190–200, 2015.

Z. Khayyat, I. F. Ilyas, A. Jindal, S. Madden, M. Ouzzani, P. Papotti, J.-A. Quiané-Ruiz, N. Tang, and S. Yin, "BigDansing," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data - SIGMOD '15*, 2015, pp. 1215–1230.

Z. Mo and Y. Li, "Research of Big Data Based on the Views of Technology and Application," *Am. J. Ind. Bus. Manag.*, vol. 05, no. 04, pp. 192–197, Apr. 2015.

Z. Sun, L. Han, W. Huang, X. Wang, X. Zeng, M. Wang, and H. Yan, "Recommender systems based on social networks," *J. Syst. Softw.*, vol. 99, pp. 109–119, 2015.

Zhou, L. Chen, and Y. He, "An unsupervised framework of exploring events on twitter: filtering, extraction and categorization," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

REPORTS

D. Abadi, M.J. Franklin, J. Gehrke, L.M. Haas, A.Y. Halevy, J.M. Hellerstein, Y.E. Ioannidis, H. V. Jagadish, D. Kossmann, S. Madden, S. Mehrotra, R. Agrawal, T. Milo, J.F. Naughton, R. Ramakrishnan, V. Markl, C. Olston, B.C. Ooi, C. Ré, D. Suciu, M. Stonebraker, T. Walter, A. Ailamaki, J. Widom, M. Balazinska, P.A. Bernstein, M.J. Carey, S. Chaudhuri, J. Dean, A. Doan, The beckman report on database research, Commun. ACM. 59 (2016) 92–99. doi:10.1145/2845915.

Executive Office of the President, Big Data: Seizing Opportunities, Preserving Values (White House, Washington, DC, 2015); http://1.usa.gov/1TSOhiG.

EY, "Big data-Changing the way businesses compete and operate," 2014.

Hewlett Packard, "Internet of things research study 2015 report," 2015

K. Thiel, T. Kötter, B. Michael, R. Silipo, and P. Winters, "Creating Usable Customer Intelligence from Social Media Data: Network Analytics meets Text Mining," 2012.

H. Kennedy, G. Moss, C. Birchall, S. Moshonas, Digital Data Analysis: Guide to tools for social media & web analytics and insights, 2013. http://2plqyp1e0nbi44cllfr7pbor.wpengine.netdna-cdn.com/files/2013/02/Digital-data-analysis-guide-to-tools.pdf

N. Preslav, RitterA., Rosenthal S., Sebastian F., Stoyanov V. "SemEval-2016 Task 4: Sentiment Analysis in Twitter",2016.

Parashar and R. Carlson, "Bridge the Gap Between Business and IT: Integrating Data into Business Workflows," *Data Informed*, 2015.

U. I. Degryse, "Here Are The New Social Risks Of The Fourth Industrial Revolution," *Social Europe*, 2016. [Online]. Available: https://www.socialeurope.eu/2016/02/here-are-the-new-social-risks-of-the-fourth-industrial-revolution/.

XU SHU, "Data cleaning and knowledge discovery in process data.", 2016.

Y. Wang, "Data Preparation for Social Network Mining and Analysis," Singapore Management University, 2014

WEBPAGES

C. T. Butts, "Package 'sna.'" 2016.

"Google Cloud NL API". https://cloud.google.com/natural-language/reference/rest/

"Google Knowledge Graph Search API" https://developers.google.com/knowledge-graph/

"Gephi Toolkit." [Online]. Available: https://gephi.org/toolkit/.

"GoogleTrends". https://www.google.com/trends/

"GraphViz"**http://www.graphviz.org/**

"DataCleaner" https://datacleaner.org/

"Freebase" https://developers.google.com/freebase/

"Indico"**https://indico.io/docs#analyze_text**

"Mongoose"**http://mongoosejs.com/**

"MPQA"**http://mpqa.cs.pitt.edu/**

"Netlytic"**https://netlytic.org/**

"NetworKit." [Online]. Available: https://networkit.iti.kit.edu/features/.

"NetworkX"**https://networkx.github.io/**

"NodeXL: Network Overview, Discovery and Exploration for Excel - Download: NodeXL Basic Excel Template 2014." .

"Pajek and Pajek-XXL Programs for Analysis and Visualization of Very Large Networks." [Online]. Available: http://mrvar.fdv.uni-lj.si/pajek/pajekman.pdf

"Pajek"**http://mrvar.fdv.uni-lj.si/pajek/**

"Querying Social Media with NodeXL." [Online]. Available: http://scalar.usc.edu/works/querying-social-media-with-nodexl/what-is-social-media?path=index.

"Semantic plugin: AlchemyAPI | Gephi blog on WordPress.com." [Online]. Available: https://gephi.wordpress.com/2010/08/10/semantic-plugin-alchemyapi/. [Accessed: 01-May-2016].

"SenticNet"**http://sentic.net/**

"Social Network Visualizer"**http://socnetv.sourceforge.net/**

"SocilMantion"**http://www.socialmention.com/**

"Stanford Parser"**http://nlp.stanford.edu/software/lex-parser.shtml**

"Sysomos"https://sysomos.com/

"Unlocking the Value of Personal Data: From Collection to Usage | World Economic Forum," 2013

"Word2vec"**http://deeplearning4j.org/word2vec#intro**

"WordNet"**https://wordnet.princeton.edu/**

"WordNetAffect"**http://wndomains.fbk.eu/wnaffect.html**

A.McCranie "Network Analysis : Dyads and Triads, Reciprocity and Transitivity", University of Michigan, ICPSR, 2015.

B.Starr "A Layman's Visual Guide To Google's Knowledge Graph Search API",SEARCH ENGINE LAND, 2016

C. Metz, "AI is transforming Google search. The rest of the web is next.," *WIRED*, 2016.

C. Parker and S. Thomson, "A recap of Davos 2016 | World Economic Forum," *World Economic Forum*. [Online]. Available: http://www.weforum.org/agenda/2016/01/a-recap-of-davos-2016.

C.Y.Lin, "Overview of Big Data Analytics", http://www.ee.columbia.edu/~cylin/course/bigdata/EECS6893-BigDataAnalytics-Lecture1.pdf, Columbia University 2015

D. Laufenberg, "Diana Laufenberg: How to learn? From mistakes," *TED Talks*, 2010. [Online]. Available: http://www.ted.com/talks/diana_laufenberg_3_ways_to_teach.html.

D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete Problems in AI Safety," Jun. 2016.

F. Ricci, L. Rokach, B. Shapira, Introduction to Recommender Systems Handbook, in: Recomm. Syst. Handb., Springer US, Boston, MA, 2011: pp. 1–35. doi:10.1007/978-0-387-85820-3_1.

J. Markoff, A Deluge of Data Shapes a New Era in Computing, New York Times. (2009). http://www.nytimes.com/2009/12/15/science/15books.html?_r=0.

M. Ferguson, http://www.ibmbigdatahub.com/blog/what-graph-analytics , 2016

M.Lesk N. US Department of Commerce, "Can you trust the Fourth Paradigm?," 2016. https://www.youtube.com/watch?v=BjH6chq70NE

MooreM., http://usatoday30.usatoday.com/news/politics/story/2012-08-01/twitter-political-index/56649678/1, 2012

J. Clark, "Google Sprints Ahead in AI Building Blocks, Leaving Rivals Wary - Bloomberg," *Bloomberg*, 2016. [Online]. Available: http://www.bloomberg.com/news/articles/2016-07-21/google-sprints-ahead-in-ai-building-blocks-leaving-rivals-wary.

O. Bowcott, "UK-US surveillance regime was unlawful 'for seven years' | UK news | The Guardian," *The Guardian*, 2015. [Online]. Available: http://www.theguardian.com/uk-news/2015/feb/06/gchq-mass-internet-surveillance-unlawful-court-nsa.

P.-W. TAM, "The Government Answers Apple in the iPhone Case - The New York Times," *The New York Times*, 2016. [Online]. Available: http://www.nytimes.com/2016/03/12/technology/the-government-answers-apple-in-the-iphone-case.html?ribbon-ad-idx=4&rref=technology&module=Ribbon&version=origin&region=Header&action=click&contentCollection=Technology&pgtype=articleover&_r=0.

Pentland A. (Sandy), "REINVENTING SOCIETY IN THE WAKE OF BIG DATA," *Edge.org*,2016.[Online].Available: https://www.edge.org/conversation/alex_sandy_pentland-reinventing-society-in-the-wake-of-big-data.

R. Silipo, I. Adae, A. Hart, B. Michael, Seven Techniques for Dimensionality Reduction, (2014). https://www.knime.org/files/knime_seventechniquesdatadimreduction.pdf (accessed April 3, 2016).

S. Mithun, C. Mulligan, G. Lapalme, N. Bouguila, S. Bergler, G. Butler, and L. Kosseim, "EXPLOITING RHETORICAL RELATIONS IN BLOG SUMMARIZATION," Concordia University, 2012.

S.G. Djorgovski "Caltech-JPL Summer School on Big Data Analytics Day 1 Lecture 1." 2015[Online].Available:https://www.youtube.com/watch?v=X6cfOnlDlZ0&list=PLTrfCHV_YfzzVxVMv1JpjbuT5gDK3WpNn.

S. Struhl, In the mood for sentiment (and counting), in: Pract. Text Anal. Interpret. Text Unstructured Data Bus. Intell., 2005: p. 272. https://www.koganpage.com/product/practical-text-analytics-9780749474010#.

T. Hannay, A new kind of science: research in the age of big data, EuroScientist. (2015). http://21ax0w3am0j23cz0qd1q1n3u.wpengine.netdna-cdn.com/wp-content/uploads/pdf/2015-06-24_Research_bigdata.pdf (accessed March 15, 2016).

UNECE, HLG Big Data Project(2014) Big Data Quality Task Team December, "A Suggested Framework for the Quality of Big Data", 2014 http://docplayer.net/1637653-A-suggested-framework-for-the-quality-of-big-data-deliverables-of-the-unece-big-data-quality-task-team-december-2014.html

APENDIX A

Source code for the application is available on the public repository Github, in the following url for everyone who desires to download the code and reproduce the results.

https://github.com/annishared/senti

The structure of the source files is the following: